UNRAVELLING THE PHYSICAL ALLOSTERIC MECHANISM OF ARAC

by

Matthew J. Brown

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the Degree of Doctor of Philosophy

Baltimore, Maryland

January 2020

Abstract

The AraC protein is the primary transcription factor regulating control of the L-arabinose operon. In this role, it is known to adopt two distinct conformations reflecting its dual roles as both a repressor (in apo-form) and inducer (in holo-form) of transcription. Despite decades of study, the overall structure of these two forms has remained unknown, as has the allosteric mechanism by which the protein converts between them. In this dissertation, I describe a series of experiments that fill in the gaps in our understanding of the AraC system, and assemble these data as well as existing hints to formulate a comprehensive physical mechanism for allosteric control of AraC. Previous work has identified the 20-residue N-terminal arm of the protein as vital for repression, but has failed to generate a method of exerting such control. I designed and created a series of homolog-based AraC chimeras, using divergent arm sequences to determine that general arm structure, rather than specific long-range contacts, is responsible for this role. The interdomain linker connecting AraC's dimerization and DNA-binding domains must drastically alter structure between the two conformations of the protein. Using an array of genetic and physical tests, I have shown that the functional nature of this structural change is of a transition between helical and non-helical states. Finally, I used a bioinformatic approach to analyze amino acid covariance among AraC homologs to locate previously hypothesized interdomain contacts stabilizing the protein's repressing form. Using the data gathered here, I posit a consistent physical molecular mechanism whereby an allosteric signal may be propagated beginning from the arabinose binding site to control the structure of AraC, allowing it to properly modulate between inducing and repressing forms.

Preface

This dissertation is the original work of Matthew J. Brown and is based on research performed as a graduate student in the laboratory of Robert Schleif at Johns Hopkins University from 2014-2019, excepting Appendix II, which was written and performed predominantly by Alexander Tischer in the laboratory of Matthew Auton at the Mayo Clinic, and written only in part by Brown. The entirety of Chapter 3 was published in Brown & Schleif 2019. Appendix II was published as presented here in Tischer, Brown, et al 2019 as an accompanying manuscript to Brown & Schleif 2019. The intent is for Chapters 2 and 4 to be submitted for future publication.

Acknowledgements

I cannot accept credit for my work without first acknowledging all the individuals and organizations that have given me the opportunity to make any accomplishments of mine possible throughout the entire process. First, I would like to thank Johns Hopkins University, its Biology Department, and the CMDB program for the opportunity and financial support during my graduate work. Thank you, of course, to the present and past members of my thesis committee: Marc Ostermeier, Karen Beemon, Carl Wu, and David Zappulla. Thank you also to the various facilities and individuals that made the many facets of my work possible: the Integrated Imaging Center, especially Michael McCaffrey and Erin Pryce for the help and allowing me to image so many gels; the NMR facility and Ananya Majumdar for help with gathering and interpreting NMR data; and Katie Tripp and the Center for Molecular Biophysics for aid preparing, gathering, and interpreting CD data.

Of course, I cannot fail to mention the Schleif Lab, and its members for immeasurable guidance and support; it would be impossible overstate my appreciation. In particular, Dr. Robert Schleif has been an incredible mentor, helping me in nearly every facet of my life at some point throughout my stay in his lab. Bob, you've been a tremendous inspiration from the very beginning of my first year. Your molecular biology course is among the most influential classes I have ever taken, and I don't believe I will ever stop singing its praises; it truly opened my eyes about how to think about science, and even pedagogy. I often think about the forks in my life's road – the potential ways my life could have gone – and the influence of your philosophies about science will forever be one of the things I am most relieved to have been exposed to. Your guidance has instilled such a direction in me, that I am comforted that I will never see science as merely data collection, the way unfortunately so many seem to, and will forever strive to understand beyond the "whats" and "wheres" of science, and into the "hows" and "whys." Beyond this, I am endlessly thankful for your seemingly endless understanding and drive to help me become a better scientist, though I know I haven't always made it easy. Additionally, thank you to Mike Rodgers, who has continued to be a tremendous help with my work and development. Mike, has never ceased impressing me with the breadth and depth of his knowledge, and the discussions between him and I – as well as between him and Bob –

have gone a long way in teaching me about the principles of various realms of science and how to think about approaching problems. Thank you also to my erstwhile labmate Ory Mayberry; your companionship in and out of the lab made the environment a pleasure, a fact highlighted by the loneliness left in your absence.

Few people would truthfully describe my path as "straightforward," and graduate school has been no exception to this rule. I faced some major setbacks – some admittedly of my own making – in the first year of my graduate studies, which put significant strain on my personal and professional development following. I hesitate to guess what my life would be like if not for the support of some amazing friends, particularly my adopted class when I was (technically) a second year. Of this class, thank you especially to Erica Boehm, Riti Gupta, Jessica Winger, and Chon Winger (who I guess should be classified as a classmate-in-law). The many, many meetings of the gang for trivia, games, escape rooms, movies, out-of-state excursions during birthday season, and so much more have frequently been the absolute highlight of my stay in Baltimore, and I look forward to continuing in the future. Over the ensuing years I have been fortunate enough to be granted a number of friendships that I am most thankful for. Thank you to Evan Hass, Kevin and Diane Lebo, Vuong Tran, Lin Quek, and Kylie Chew for starting our (more or less) weekly meetings of the CMDB game night. Thank you also to all the further members who joined into the tradition over the years, including but not limited to Andrew Gale, Ann Mendoza, Rashi Turnianski, Nick Wilkerson, and Nathan Roach. Thank you especially to Jessie Kirshner, for being among my closest friends during our time together at Hopkins, and being- well, "supportive" isn't always exactly the right word, but "sarcastically unsupportive" is just as good.

Thank you also do my wonderful Boston Terrier, Parker. I'm holding out hope that someday he'll be reading this. Okay, so I'll *probably* have to read it to him, but I think we can get there; he currently at least knows the words "sit," "down," "paw," "bed," "donut," "cheese," "ball," and "milk," as well as his own name, so I think we're getting there. Truthfully, though, he's been an incalculable comfort in these last 5 years, and I can't count the number of times his ridiculous antics have brightened an otherwise lousy day.

Although it may not seem like it anymore, life did exist before grad school, and I need to thank the people who helped me get there. Thank you to the friends who stuck with me to this point and beyond, especially Shane and Alicia Dewar, and Zack Carvalho. Though it sounds clichéd to say, I truly do not remember a time when I was not interested in pursuing science, which I can only credit to my wonderful family. In my youth, my mother the science educator, and my father the chemical engineer endowed in me a tremendous interest in all things science. In retrospect, this is especially impressive given the many issues I brought to the table; I remain astounded that my parents and my sister were able to do such a excellent job fostering my intellectual curiosity while simultaneously having to deal with me assaulting elementary school teachers and setting tables on fire at school (okay, but that only happened that one time!) as well as frequent refusals to actually do my school work. I know I haven't always made it easy, but words cannot state how grateful I am for all their patience and hard work raising me. I can only say the words that I have to believe every parent wants to hear: all those times you told me I would understand when I'm older? You were right. …But also kind of wrong about that; I don't know if I'll ever understand how you could have managed to handle all my… everything. In mentioning my upbringing, I could not possibly leave out my sister, Heather; I will probably never appreciate how important you were in my early development, but I truly do appreciate all the ongoing help and support you have continued to give me over the years. I know I don't always seem appreciative; it is impossible not to feel like the baby every time I'm around, but that is only because you are so smart and capable that, even as this dissertation marks me gaining a doctorate degree, I can't help feeling dwarfed by your shadow. To my family: in every sense, I owe any success I have to you.

Table of contents

Abstract ii

Preface iii

Acknowledgements iv

Table of contents vii

List of Tables xi

List of Figures xii

Chapter 1: Allosteric control in proteins and the L-arabinose operon 1

Understanding the mechanisms of allosteric control 1

Allosteric control in transcription factors 3

AraC and the arabinose operon 6

References 11

Chapter 2: AraC orthologs reveal previously unknown intra-domain contacts 17

Abstract 17

Introduction 17

Results 22

In vivo behavior of AraC orthologs 22

Discussion 23

Materials and Methods 26

Plasmids and strains 26

Arabinose isomerase assay 26
References 27
Chapter 3: Helicity in the Interdomain Linker in AraC 28
Abstract 28
Introduction 28
Results 32
Helix stabilizing and destabilizing mutants – *in vivo* properties 32
Excluding a potential artifact 33
In vitro properties of the helix stabilizing and destabilizing variants 34
Helix propagation into the linker sequence 36
Discussion 38
Materials and Methods 42
Plasmids and strains 42
DNA used in binding measurements 42
Protein used in DNA binding measurements 43
Arabinose isomerase assay 43
DNA dissociation and electrophoretic mobility shift assay 43
Circular dichroism (CD) measurements and sample preparation 44
References 44
Chapter 4: Residue covariance identifies interdomain contacts 47
Abstract 47
Introduction 47

Results 50
Generation of input dataset 50
Prediction of functional contacts 51
Effect of variation between predicted residue pairs 54
Discussion 56
Materials and Methods 59
Identification of AraC homologs 59
Analysis of covariance using GREMLIN software 59
Generation of AraC mutants 59
p_{BAD}-GFP fluorescence assay 60
References 60
Chapter 5: Conclusion/Perspectives – Where we have come to with AraC 63
Abstract 63
Discussion 63
Step one: Arm folds over arabinose 64
Step two: Linker transitions out of helical state 67
Step three: Interdomain contacts stabilize new conformation 67
References 70
Appendix I: Python scripts 71
MSA-1line.py 71
ResRand.py 71
AllContacts.py 73

Appendix II: Arabinose Alters Both Local and Distal H-D Exchange Rates in the *Escherichia coli* AraC Transcriptional Regulator..... 76

Abstract..... 78

Introduction..... 80

Methods..... 82

Protein Expression and Purification..... 82

Hydrogen Deuterium Exchange Mass Spectrometry (HXMS)..... 82

Results..... 83

Hydrogen-deuterium exchange of the dimerization domain..... 83

Simultaneous stabilization and relaxation of the arabinose binding barrel..... 85

Allosteric relaxation of the C-terminal interdomain linker..... 86

Hydrogen-deuterium exchange of full-length AraC confirms observations in the dimerization domain and reveals allosteric stabilization of the DNA binding domain..... 87

Discussion..... 88

Acknowledgements..... 91

References..... 91

Curriculum Vitae..... 107

List of Tables

Table 3.1. *In vivo* activity of AraC linker mutants on $arap_{BAD}$ in units per cell……...31

Table 3.2. Hypoinducibility of the 6A linker is dominant to wild type AraC……..32

Table 3.3. Sequence layout of linker-capped peptides…….35

Table 4.1. GREMLIN-predicted DD-DBD contact locations …….54

List of Figures

Figure 1.1. Regulatory modes of AraC……7

Figure 1.2. Structure of AraC homodimer……...9

Figure 2.1. Possible mechanisms for arm control of repression……17

Figure 2.2. *In vivo* expression of arm mutants in *E. coli* AraC……..21

Figure 3.1. Overview of AraC……..……...28

Figure 3.2. DNA electrophoretic mobility shift measuring the dissociation kinetics of AraC protein from I_1-I_1 site in the presence of arabinose and 300 mM KCl……...33

Figure 3.3. Dissociation of I_1-I_1 from AraC linker variants as a function of time…….33

Figure 3.4. CD spectra of helix/linker peptides…….36

Figure 3.5. The positions of repression-disrupting mutations……39

Figure 4.1. Example of identification of nucleotide interaction by covariance…….48

Figure 4.2. Enrichment of nearby residues in GREMLIN-predicted pairs……51

Figure 4.3. GREMLIN-predicted locations of possible inter-domain contacts in AraC……...53

Figure 4.4. Potential model of DBD-DD interaction…….56

Figure 5.1. Mechanism of AraC…….63

Figure 5.2. Overlay of apo and holo dimerization domain……64

Figure 5.3. DBD-DD interdomain contacts……...67

Figure 5.4. DNA-binding domain/dimerization domain interface……68

Chapter 1: Allosteric control in proteins and the L-arabinose operon

Understanding the mechanisms of allosteric control

In the current state of biological science, the relevant scale for notable understanding is, at broadest, molecular. As is the case for any science, the history of biological study has been an ever-sharpening focus progressing towards understanding of mechanisms and basic principles. Developing initially from broad physiological and ecological abstractions, in several areas, the field has currently reached the point of approaching the fundamental molecular interactions behind biological processes. Both the goal and ultimate test of scientific understanding is engineering, in this case, the capacity to predict, tune, and ultimately design biological molecules *de novo*. Thus, even beyond studying intermolecular interactions, intramolecular processes seem to hold the key to fundamental understanding of biology. Allostery denotes the critical mechanisms behind some of these intramolecular processes, and the study thereof should elucidate some of the physical principles governing proteins' behavior, and – at a fundamental level – just how many important proteins and other biological molecules operate. Unfortunately, at the time of this writing, few allosteric systems have yet been studied sufficiently to build much of a vocabulary of mechanistic principles.

Much of the initial attempts to formalize mechanisms of allosteric conformational change come from work done by Monod, Wyman, and Changeux(Monod, Changeux et al. 1963), ultimately described by their concerted (MWC) model(Monod, Wyman et al. 1965). Lacking the ability to easily asses comprehensive protein structures, they created a phenomenological model, founded upon 6 points: (1) an allosteric protein is an oligomer composed of equivalent subunits; (2) each subunit is capable of binding exactly one ligand; (3) contact between subunits constrains their conformations; (4) at any time all the allosteric protein exists in one of two alternative, interchangeable states, distinguished by differential ligand affinity; (5) the ligand affinity of any site is thus altered by transition between these states; (6) molecular symmetry is conserved upon transition between states(Monod, Wyman et al. 1965). Thus, the key of the

MWC model is the notion that allosteric proteins freely interchange between two states, independent of ligand, and that ligand binding alters the thermal equilibrium of these states, biasing the protein towards one state or the other.

An alternative model was, shortly after, proposed by Koshland, Némethy and Filmer(Koshland, Nemethy et al. 1966), ultimately taking a more localized view of allostery. This sequential (KNF) model treats the subunits of an allosteric protein more independently, hypothesizing that protein subunits initially react to ligand binding via an induced fit method – ligand binding produces a local conformational change(Koshland, Nemethy et al. 1966). Intramolecular contacts would subsequently produce changes in nearby subunits. In such a manner, this ensuing restructuring would produce successive effects among further adjacent subunits. Thus, the KNF model describes a system whereby allosteric signal is produced by sequential structural changes within the protein.

Both the MWC and KNF models describe systems whereby protein subunits communicate across interfaces in some manner. Each model has been applied in various systems(Yu, Koshland 2001, Yifrach, MacKinnon 2002, Velyvis, Schachman et al. 2009, Purohit, Mitra et al. 2007), making apparent that different allosteric systems adequately be described by somewhat different sets of system assumptions. The notion that different proteins can operate by different means is not in question, however, and indeed it is the various differences that must be understood in order to facilitate eventual predictive models. The MWC and KNF models themselves are abstractions, at best describing overall classes of complicated interactions, and only useful as *ex-post-facto* fits of observed data. To meaningfully deepen our understanding of allostery, the goal must be to characterize the precise nature of interactions between subunits and the consequences of these interactions.

Both models may appear to describe general approaches to allosteric control, but in reality, they are more accurately described as statistical methods of fitting observed data. Rather than describing physical phenomena, these general models represent abstractions – generalized sets of rules that fail to actually characterize what is happening within the molecule. While investigators studies have fit their data into one

model or the other, this *ex-post-facto* approach offers little, if any, predictive power. To meaningfully advance our understanding of molecular mechanisms, it is imperative that research be more focused on identifiable systems arising out of basic principles. The experimental work described in this dissertation represents an attempt to identify the relevant molecular interactions and their consequences.

Allosteric control in transcription factors

Thus far, much of the research into allosteric mechanisms of proteins have been carried out in enzymes. This is not surprising, as a direct biochemical readout produced by such enzymes provides a direct and relatively unambiguous means of report, allowing for detailed kinetics analysis. By comparison, study of allostery in transcription factors as lagged somewhat, due in no small part to the curious observation that allosteric signals in transcription factors almost always must be transmitted over much longer distances than in enzymes. Several of the historically-popular transcription factors have been somewhat characterized, however.

The lactose operon repressor LacI has long been the bellwether for studies of transcription factors, including allostery research. LacI represses the *lac* operon in the absence of allolactose, binding over the operator and blocking RNA polymerase from binding(Jacob, Monod 1961, Gilbert, Maxam 1973). Intracellular lactose, when available, is converted into allolactose, which binds LacI, releasing it from its repressive binding, allowing transcription of the *lac* genes(Jacob, Monod 1961). The LacI protein itself is a homotetramer, with each subunit consisting of an N-terminal DNA binding domain and a regulatory domain, connected via an 18-residue linker sequence(Lewis, Chang et al. 1996). This linker is perhaps the most vital portion for allosteric control of LacI, containing a central "hinge helix" flanked by unstructured regions at both ends(Lewis, Chang et al. 1996). This hinge helix is often cited as a primary region of structural alteration between the repressing and non-repressing conformations of LacI, at the core of how the protein's DNA-binding affinity is altered when allolactose is bound (Lewis, Chang et al. 1996).

Aside from the hinge helix, other, disparate, regions within both the regulatory and DNA-binding domains have been implicated(Muller-Hartmann, Muller-Hill 1996, Daber, Sochor et al. 2011, Wilson, Zhan et al. 2007, Taraban, Zhan et al. 2008) in allosteric control, but the nature of their connectivity appears, at first glance, unclear. However, an understanding of the molecular mechanism may be achieved by acknowledging the lac repressor tetramer as a pair of dimers, with a single dimer binding a single DNA binding site. As suggested by the name, it has previously been suggested that the hinge helix region rotates the surrounding regions with respect to each other in a hinge-like motion when the regulatory domain binds allolactose(Lewis, Chang et al. 1996b), but a more careful examination of the structure reveals a different, more likely mechanism. The hinge helices of paired subunits within a dimer lie next to each other, contacting the minor groove within the center of a DNA binding site. This positioning allows the two associated DNA-binding domains to contact the flanking major grooves of the site. Ligand binding to its site on LacI folds the regulatory domain slightly, causing the hinge helices to move relative to one another; this has the effect of both disrupting their fit in the DNA minor groove and altering their local environment (the helical structure of one is no longer stabilized by contact with the other). This, in turn, leads to the helices no longer stabilizing the locations of the DNA-binding domains, resulting in lowered DNA affinity. In this manner, the pathway can be stated as progressing from (1) the ligand binding site, to (2) large-scale rearrangement of the regulatory domain, to (3) hinge helix, and finally (4) DNA-binding domain.

Another heavily-studied protein, cyclic AMP receptor protein (CRP) – also involved in regulation of the *lac* operon (among many other genes)(Busby, Ebright 1999) seems to operate by a superficially similar but distinct allosteric mechanism. CRP binds to its target DNA only in the presence of cyclic AMP. Another oligomer, CRP exists as a dimer(McKay, Weber et al. 1982). Like LacI, CRP comprises two domains (a DNA-binding domain, and a cAMP-binding domain) connected by a partially helical linker. This linker, also called a hinge (not to be confused with the hinge helix of LacI) is the site of dimerization between CRP protomers. Binding of CRP to its ligand, cAMP, alters the structure of these hinge linkers, extending the helices. Compared to LacI, this mechanism seems to be somewhat more immediately apparent, as cAMP binds adjacent to these linkers, directly contacting and stabilizing the helicity of the region. This ultimately

causes a shift in the DNA-binding domains, reorienting them into positions more accessible for DNA-binding, drastically increasing DNA-binding affinity.

Unfortunately, the nature of this conformational shift is somewhat ambiguous. Stepping back, one can easily hypothesize three distinct structural methods of abolishing DNA binding: first a protein's DNA binding contacts might be structurally altered to no longer recognize/fit the DNA helix; second, the contacts could be buried or occluded – keeping the proper local structure but preventing access to DNA; third, the tethering of separate DNA-binding domains could be loosened, increasing the flexibility and causing entropic inhibition as the different subunits lose the statistical probability of simultaneously existing in the proper relative locations. In the case of CRP, it is not immediately clear which of these means of control are at play in the non-binding apo-form. Structures derived from X-ray crystallography data show regions near the DNA-binding faces of the proteins contacting each other, obscuring them from binding DNA(Sharma, Yu et al. 2009). Alternatively, NMR imaging of CRP seems to suggest the flexible tether model, as it notably lacks the apparent contacts seen in the X-ray structures(Popovych, Tzeng et al. 2009), although this may be a consequence of a spin-label interfering with such interactions. The ambiguity is somewhat resolved by later physical experiments whose most parsimonious explanation echoed the interdomain contact model suggested by the X-ray crystal structures(Yu, Maillard et al. 2012).

The proposed mechanisms of LacI and CRP illustrate the importance of building a repertoire of allosteric pathways. It could be easy to lump both systems as being dimers altering overall arrangement of subunits by shifting linker structures, but this is a rather superficial characterization. While using older statistical mechanisms, analysis of data within the various frameworks could not differentiate between the two, the actual mechanisms are distinct. The case of CRP, also, highlights an important consideration of physical allosteric studies: the actual nature of the end-states. CRP's DNA-binding domains could plausibly have been controlled enthalpically by favorable association of nearby regions, or entropically by flexible tethering maximizing their degrees of freedom, and the answer is vital to understanding the nature of the transition. Statistical models merely note that "something" changes, but the actual physical nature of that

change could potentially be a number of drastically different states. If the ultimate goal of scientific investigation is engineering, the field is still reverse-engineering the rules, and building a repertoire – finding the biological equivalents of the wheel, lever, and inclined plane – is necessary.

AraC and the arabinose operon

Metabolism of the plant sugar arabinose in *E. coli* – as well as many other bacterial species –is regulated by the L-arabinose operon, and the process is regulated by the transcription factor AraC(Sheppard, Englesberg 1967). The *araA*, *araB*, and *araD* genes within this operon encode the proteins responsible for this metabolic pathway, by which L-arabinose is converted into D-xylulose phosphate for consumption in central metabolism(GROSS, ENGLESBERG 1959). The AraC protein is responsible for regulation of the entire operon, including its own structural gene, *araC*, also located within this region. As shown in Figure 1, it accomplishes this by binding to alternative DNA sites in the upstream regulatory region of the operon.

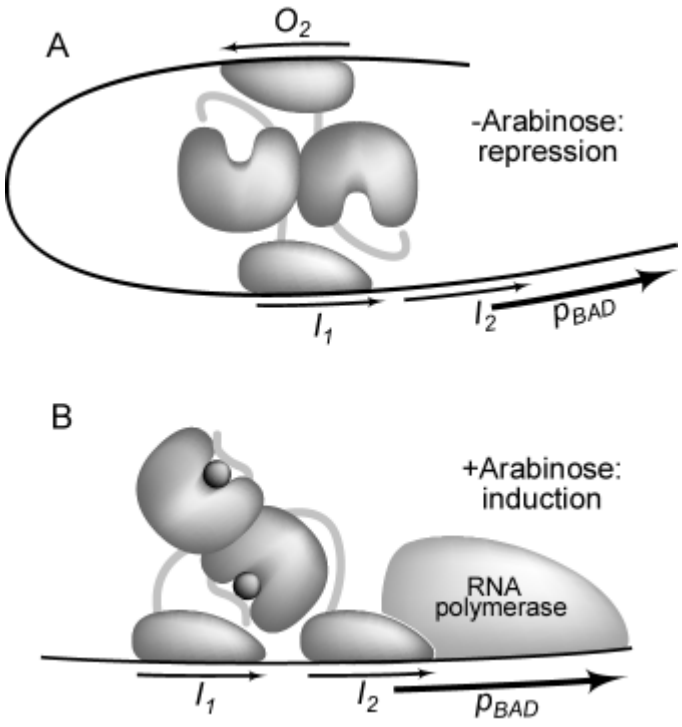

Figure 1. Regulatory modes of AraC The DNA binding sites in the *ara* regulatory region contacted by AraC in the (**A**) absence and (**B**) presence of arabinose. The p_{BAD} is the RNA polymerase binding site.

AraC regulates the arabinose operon by both repressive and inductive means(Englesberg, Irr et al. 1965, Sheppard, Englesberg 1967). This process is itself regulated by the presence of intracellular arabinose(Gross, Englesberg 1959, Steffen, Schleif 1977). AraC represses the genes of the operon when arabinose is absent, but activates transcription when arabinose is available and bound to the protein. AraC exists as a homodimer(Steffen, Schleif 1977), and – when no arabinose is present – binds to two somewhat distant DNA sites, looping DNA(Dunn, Hahn et al. 1984, Lee, Schleif 1989), repressing expression from the p_{BAD} promoter as this structure physically occludes RNA polymerase from the promoter(Hahn, Schleif 1983). These two sites are designated I_1, which is directly upstream from the core promoter, and O_2, 210 bp further upstream (Figure 1A). When arabinose is present in the cell, and binds to AraC, the protein shifts its DNA-binding pattern to the directly adjacent I_1 and I_2 sites(Carra, Schleif 1993, Dunn, Hahn et al. 1984) (Figure 1B). In this configuration, not only does the loop-based inhibition of transcription cease, but AraC

actively stimulates productive binding of RNA polymerase to the p_{BAD} promoter, both stabilizing the RNAP-DNA binding, and accelerating the transition from the from the closed transcription complex to an open complex(Zhang, Reeder et al. 1996) (though the mechanisms behind this this protein-protein interaction remain largely unclear). Thus, AraC regulates the arabinose operon in two ways: (1) by looping/unlooping DNA, and (2) by facilitating the action of RNA polymerase at the promoter.

The gross structural overview an AraC dimer is illustrated in Figure 2. Each monomeric subunit comprises two domains, the dimerization domain and the DNA-binding domain (DBD), connected by a short (8-residue) interdomain linker. The first 20 residues of the protein constitute the N-terminal arm, which immediately precedes the dimerization domain. The dimerization domain additionally contains the arabinose-binding pocket, located directly adjacent to the arm. Also near this region is the dimerization interface, which consists of an antiparallel coiled-coil leading directly into the interdomain linker, and the DNA-binding domain beyond.

The DNA binding patterns of apo- (repressing) and holo- (inducing) AraC necessitate significant structural changes within the protein, and it is the nature of these changes, and the mechanisms driving them, that I aim to unravel throughout this dissertation. While one might expect X-ray crystallography to be the most effective means for identifying much of the relevant structural differences between these forms, such efforts have been largely unsuccessful, as decades of attempts have consistently failed to crystalize the full-length protein. The individual domains, however have been separately solved(Soisson, MacDougall-Shackleton et al. 1997, Rodgers, Schleif 2009a, Weldon, Rodgers et al. 2007). Comparing the apo- and holo- structures of the dimerization domain, previous research has noted that the N-terminal arm changes structure from a “free” position protruding away from the ligand-binding site (represented in the cartoon layer of Figure 2), to binding over arabinose when the sugar is bound (seen in the crystal structure overlay of Figure 2)(Weldon, Rodgers et al. 2007, Soisson, MacDougall-Shackleton et al. 1997). The DBDs must alter positions – relative to the dimerization domains – to adopt different binding patterns, but without full-length

crystal structures, the overall conformations of the protein dimer remain unknown, and thus, the larger-scale structural changes between the apo- and holo- states cannot be satisfactorily determined via such methods.

Thus, the overarching question to be answered is “how?” How does AraC alter its structure in response to arabinose binding? What areas of the protein are responsible for the structural shift? Which

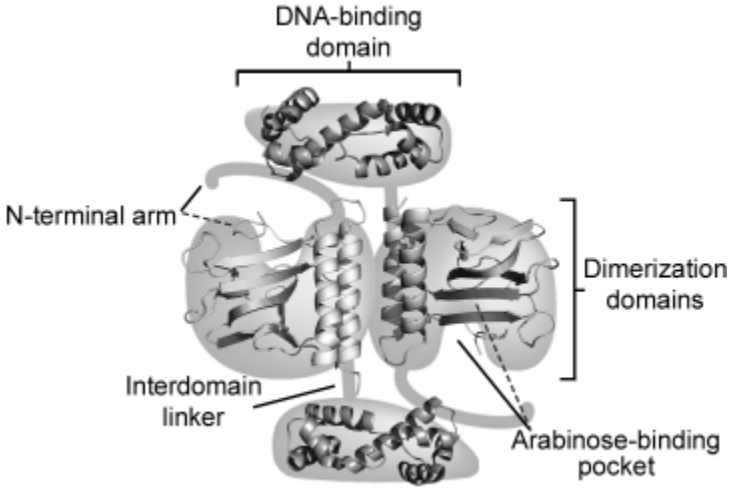

Figure 2. Structure of AraC homodimer. Domain crystal structures (Dimerization domains PDB: 2ARC and DNA-binding domain PDB: 1XJA (2x)) overplayed on cartoon representation. Dashed lines refer to regions on crystal structure where overlay is not perfectly aligned with cartoon. Relative positions of domains is a loose prediction, and has not been experimentally determined

regions interact with each other? What is the nature of the allosteric signal, and how does it propagate from the site of ligand binding to the DBDs? Numerous lines of experimentation have provided hints to various aspects this conformational shift’s nature, and the experiments reported here have allowed me to assemble these into a plausible allosteric model.

Since the N-terminal arm seems to be located immediately adjacent to the arabinose-binding pocket, and it alters structure between the two forms and binds directly over the ligand, it is an attractive candidate for a putative site of signal initiation. As shown in Figures 1 and 2, the apo arm structure likely places it near the DBD of the opposite subunit, leading to a plausible line communication pathway whereby it directly interacts with the DBD in this state, but releases when arabinose is added. However, though a great many arm mutations drastically hinder proper repression(Ross, Gryczynski et al. 2003, Dirla, Chien et al. 2009),

no similar mutations have been found in the DBD, nor any compensating mutations, despite extensive testing. Additionally, NMR of constructs with the arm sequence attached to the DBD, to maintain proper local concentration, have similarly failed to identify association between the two regions. However, there is evidence that the DBD interacts with the dimerization domain *somewhere*, as the DBD alone shows altered DNA-binding affinity compared to full-length AraC(Cole, Schleif 2012), and a DNA-assisted binding assay demonstrated some level of weak interaction between the two separate domains(Frato, Schleif 2009).

The interdomain linker is another attractive candidate for a site of allosteric regulation. As suggested by Figure 1, it seems that the linker must alter structure significantly between the repressing and inducing structures in order to bind alternative DNA sites. Indeed, past experiments have shown that arabinose binding triggers some manner of change in the linker. Fluorescence anisotropy – in experiments labeling the end of the linker of dimerization domain(Malaga, Mayberry et al. 2016), or the DNA-binding domain of full-length AraC(Mayberry, unpublished) – show an increase in flexibility upon addition of arabinose *in vitro*. This recapitulates hints suggested by some of the earlier work characterizing DNA looping in the arabinose operon; moving the O_2 site, by either inversion or shifting 5 bp to the opposite 3-dimensional face of the DNA, eliminates repression, but inversion of the I_2 site does not affect induction in the same way(Seabold, Schleif 1998, Carra, Schleif 1993). Thus, we conclude that the linker adopts a rigid, inflexible structure in the repressing form, but represents more of a flexible tether in the inducing state. How could this apparent adjustment in flexibility be attained? A hint to this question comes from examining a screen of linker mutations. A large majority of partial (1-3-residue) linker substitutions display no discernable phenotype, but introductions of proline residues nearly all inhibit repression(Malaga, Mayberry et al. 2016). This suggests a possible mechanism whereby linker rigidity may be established and maintained by modulation of linker secondary structure, specifically α-helical structure. This mechanism will be further tested in Chapter 3.

As mentioned above, allosteric signals in transcription factors must communicate across long molecular distances. The Lac repressor and CRP systems show some “clever” ways this can be

accomplished. These two proteins, as well as what we know about AraC, suggest a theme of large-scale rearrangement shifting the positions of DNA-binding domains. Compared to the alternative of directly altering the DNA-binding interfaces, this class of mechanism effectively shortens the distance the allosteric signal needs to move, as it needs only transmit as far as an interdomain linker. For LacI, this seems to be accomplished by a large-scale change loosening the packing of a linker region; In CRP, it appears to be accomplished by altering the linker structure directly; In the case of AraC, there is already strong evidence that a cascade of interactions breaks the helical structure of a linker, allowing the DNA-binding domains to shift into new positions to bind alternative sites. This type of mechanism is promising for the eventual goal of protein engineering, as it could provide a means of predictably altering the structure of an allosteric protein that already appears widely used in nature.

References

BERG, O. and VON HIPPEL, P.H., 1985. Diffusion-controlled macromolecular interactions. *Annu Rev Biophys Biophys Chem,* **14**, pp. 131-160.

BRAHMS, S. and BRAHMS, J., 1980. Determination of protein secondary structure in solution by vacuum ultraviolet circular dichroism. *Journal of Molecular Biology,* **138**(2), pp. 149-178.

BUSBY, S. and EBRIGHT, R.H., 1999. Transcription activation by catabolite activator protein (CAP). *Journal of Molecular Biology,* **293**(2), pp. 199-213.

BUSTOS, S.A. and SCHLEIF, R.F., 1993. Functional domains of the AraC protein. *Proceedings of the National Academy of Sciences of the United States of America,* **90**(12), pp. 5638-5642.

CARRA, J.H. and SCHLEIF, R.F., 1993. Variation of half-site organization and DNA looping by AraC protein. *The EMBO journal,* **12**(1), pp. 35-44.

COLE, S.D. and SCHLEIF, R., 2012. A new and unexpected domain-domain interaction in the AraC protein. *Proteins: Structure, Function, and Bioinformatics,* **80**(5), pp. 1465-1475.

DABER, R., SOCHOR, M.A. and LEWIS, M., 2011. Thermodynamic Analysis of Mutant lac Repressors. *Journal of Molecular Biology; The Operon Model and its Impact on Modern Molecular Biology,* **409**(1), pp. 76-87.

DIRLA, S., CHIEN, J.Y. and SCHLEIF, R., 2009. Constitutive mutations in the *Escherichia coli* AraC protein. *Journal of Bacteriology,* **191**(8), pp. 2668-2674.

DUNN, T.M., HAHN, S., OGDEN, S. and SCHLEIF, R.F., 1984. An operator at -280 base pairs that is required for repression of *araBAD* operon promoter: addition of DNA helical turns between the operator and promoter cyclically hinders repression. *Proceedings of the National Academy of Sciences of the United States of America,* **81**(16), pp. 5017-5020.

EDDY, S.R. and DURBIN, R., 1994. RNA sequence analysis using covariance models. *Nucleic acids research,* **22**(11), pp. 2079-2088.

ENGLESBERG, E., IRR, J., POWER, J. and LEE, N., 1965. Positive control of enzyme synthesis by gene C in the L-arabinose system. *Journal of Bacteriology,* **90**(4), pp. 946-957.

EUSTANCE, R.J., BUSTOS, S.A. and SCHLEIF, R.F., 1994. Reaching out. Locating and lengthening the interdomain linker in AraC protein. *Journal of Molecular Biology,* **242**(4), pp. 330-338.

FRATO, K.E. and SCHLEIF, R.F., 2009. A DNA-assisted binding assay for weak protein-protein interactions. *Journal of Molecular Biology,* **394**(5), pp. 805-814.

GALLEGOS, M.T., SCHLEIF, R., BAIROCH, A., HOFMANN, K. and RAMOS, J.L., 1997. Arac/XylS family of transcriptional regulators. *Microbiology and molecular biology reviews : MMBR,* **61**(4), pp. 393-410.

GHOSH, M. and SCHLEIF, R.F., 2001. Biophysical evidence of arm-domain interactions in AraC. *Analytical Biochemistry,* **295**(1), pp. 107-112.

GILBERT, W. and MAXAM, A., 1973. lac Operator Structure and Repressor Interaction Sites. *Proc.Nat.Acad.Sci.USA,* **70**, pp. 3581.

GROSS, J. and ENGLESBERG, E., 1959. Determination of the order of mutational sites governing L-arabinose utilization in Escherichia coli B/r bv transduction with phage Plbt. *Virology,* **9**, pp. 314-331.

GUTFREUND, H., 1995. *Kinetics for the Life Sciences.* Cambridge, UK: Cambridge University Press.

HAHN, S., DUNN, T. and SCHLEIF, R., 1984. Upstream repression and CRP stimulation of the *Escherichia coli* L-arabinose operon. *Journal of Molecular Biology,* **180**(1), pp. 61-72.

HAHN, S. and SCHLEIF, R., 1983. *In vivo* regulation of the *Escherichia coli araC* promoter. *Journal of Bacteriology,* **155**(2), pp. 593-600.

HENDRICKSON, W. and SCHLEIF, R., 1985. A dimer of AraC protein contacts three adjacent major groove regions of the *araI* DNA site. *Proceedings of the National Academy of Sciences of the United States of America,* **82**(10), pp. 3129-3133.

HENDRICKSON, W. and SCHLEIF, R.F., 1984. Regulation of the *Escherichia coli* L-arabinose operon studied by gel electrophoresis DNA binding assay. *Journal of Molecular Biology,* **178**(3), pp. 611-628.

JACOB, F. and MONOD, J., 1961. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology,* **3**(3), pp. 318-356.

KAMISETTY, H., OVCHINNIKOV, S. and BAKER, D., 2013. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci USA,* **110**(39), pp. 15674.

KIM, M.K. and KANG, Y.K., 1999. Positional preference of proline in α-helices. *Protein Science,* **8**(7), pp. 1492-1499.

KOSHLAND, D.E., Jr., NEMETHY, G. and FILMER, D., 1966. Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry,* **5**(1), pp. 365-85.

KUKIC, P., CAMILLONI, C., CAVALLI, A. and VENDRUSCOLO, M., 2014. Determination of the Individual Roles of the Linker Residues in the Interdomain Motions of Calmodulin Using NMR Chemical Shifts. *Journal of Molecular Biology,* **426**(8), pp. 1826-1838.

LEE, D. and SCHLEIF, R., 1989. *In vitro* DNA Loops in *araCBAD*: Size Limits and Helical Repeat. *Proc.Nat.Acad.Sci.USA,* **86**, pp. 476-480.

LEWIS, M., CHANG, G., HORTON, N.C., KERCHER, M.A., PACE, H.C., SCHUMACHER, M.A., BRENNAN, R.G. and LU, P., 1996. Crystal structure of the lactose operon represser and its complexes with DNA and inducer. *Science,* **271**(5253), pp. 1247-1254.

LOBELL, R.B. and SCHLEIF, R.F., 1990. DNA looping and unlooping by AraC protein. *Science (New York, N.Y.),* **250**(4980), pp. 528-532.

MA, B., TSAI, C., HALILOĞLU, T. and NUSSINOV, R., 2011. Dynamic Allostery: Linkers Are Not Merely Flexible. *Structure,* **19**(7), pp. 907-917.

MALAGA, F., MAYBERRY, O., PARK, D., RODGERS, M.E., TOPTYGIN, D. and SCHLEIF, R.F., 2016. A genetic and physical study of the interdomain linker of E. Coli AraC protein-a trans-subunit communication pathway. *Proteins: Structure, Function, and Bioinformatics,* **84**, pp. 448--460.

MARQUSEE, S., ROBBINS, V.H. and BALDWIN, R.L., 1989. Unusually stable helix formation in short alanine-based peptides. *Proceedings of the National Academy of Sciences of the United States of America,* **86**(14), pp. 5286-5290.

MARTIN, K., HUO, L. and SCHLEIF, R.F., 1986. The DNA loop model for *ara* repression: AraC protein occupies the proposed loop sites *in vivo* and repression-negative mutations lie in these same sites. *Proceedings of the National Academy of Sciences of the United States of America,* **83**(11), pp. 3654-3658.

MCKAY, D., WEBER, I. and STEITZ, T., 1982. Structure of Catabolite Gene Activator Protein at 2.9 Â Resolution. *J.Biol.Chem.,* **257**, pp. 9518-9524.

MONOD, J., CHANGEUX, J.P. and JACOB, F., 1963. Allosteric proteins and cellular control systems. *Journal of Molecular Biology,* **6**, pp. 306-329.

MONOD, J., WYMAN, J. and CHANGEUX, J.P., 1965. On the Nature of Allosteric Transitions: a Plausible Model. *Journal of Molecular Biology,* **12**, pp. 88-118.

MULLER-HARTMANN, H. and MULLER-HILL, B., 1996. The side-chain of the amino acid residue in position 110 of the Lac repressor influences its allosteric equilibrium. *Journal of Molecular Biology,* **257**(3), pp. 473-478.

OVCHINNIKOV, S., KAMISETTY, H. and BAKER, D., 2014. Robust and accurate prediction of residueâ€“residue interactions across protein interfaces using evolutionary information. *eLife,* **3**, pp. e02030.

PACE, C.N. and SCHOLTZ, J.M., 1998. A Helix Propensity Scale Based on Experimental Studies of Peptides and Proteins. *Biophysical journal,* **75**(1), pp. 422-427.

PARKER, A. and GOTTESMAN, S., 2016. Small RNA Regulation of TolC, the Outer Membrane Component of Bacterial Multidrug Transporters. *J. Bacteriol.,* **198**(7), pp. 1101-1113.

POPOVYCH, N., TZENG, S.R., TONELLI, M., EBRIGHT, R.H. and KALODIMOS, C.G., 2009. Structural basis for cAMP-mediated allosteric control of the catabolite activator protein. *Proceedings of the National Academy of Sciences of the United States of America,* **106**(17), pp. 6927-6932.

PRESTA, L.G. and ROSE, G.D., 1988. Helix signals in proteins. *Science,* **240**(4859), pp. 1632.

PUROHIT, P., MITRA, A. and AUERBACH, A., 2007. A stepwise mechanism for acetylcholine receptor channel gating. *Nature,* **446**(7138), pp. 930-933.

REED, W.L. and SCHLEIF, R.F., 1999. Hemiplegic mutations in AraC protein. *Journal of Molecular Biology,* **294**(2), pp. 417-425.

RODGERS, M.E., HOLDER, N.D., DIRLA, S. and SCHLEIF, R., 2009. Functional modes of the regulatory arm of AraC. *Proteins,* **74**(1), pp. 81-91.

RODGERS, M.E. and SCHLEIF, R., 2009. Solution structure of the DNA binding domain of AraC protein. *Proteins,* **77**(1), pp. 202-208.

RODGERS, M.E. and SCHLEIF, R., 2012. Heterodimers reveal that two arabinose molecules are required for the normal arabinose response of AraC. *Biochemistry,* **51**(41), pp. 8085-8091.

ROSS, J.J., GRYCZYNSKI, U. and SCHLEIF, R., 2003. Mutational analysis of residue roles in AraC function. *Journal of Molecular Biology,* **328**(1), pp. 85-93.

SCHLEIF, R.F. and WENSINK, P.C., 1981. *Practical methods in molecular biology.* New York: Springer-Verlag.

SEABOLD, R.R. and SCHLEIF, R.F., 1998. Apo-AraC actively seeks to loop. *Journal of Molecular Biology,* **278**(3), pp. 529-538.

SEEDORFF, J. and SCHLEIF, R., 2011. Active role of the interdomain linker of AraC. *Journal of Bacteriology,* **193**(20), pp. 5737-5746.

SHARMA, H., YU, S., KONG, J., WANG, J. and STEITZ, T.A., 2009. Structure of apo-CAP reveals that large conformational changes are necessary for DNA binding. *Proceedings of the National Academy of Sciences,* **106**(39), pp. 16604-16609.

SHEPPARD, D.E. and ENGLESBERG, E., 1967. Further evidence for positive control of the L-arabinose system by gene *araC*. *Journal of Molecular Biology,* **25**, pp. 443-454.

SOISSON, S.M., MACDOUGALL-SHACKLETON, B., SCHLEIF, R. and WOLBERGER, C., 1997. The 1.6 A crystal structure of the AraC sugar-binding and dimerization domain complexed with D-fucose. *Journal of Molecular Biology,* **273**(1), pp. 226-237.

STEFFEN, D. and SCHLEIF, R., 1977. *In vitro* construction of plasmids which result in overproduction of the protein product of the *araC* gene of *Escherichia coli*. *Molecular & general genetics : MGG,* **157**(3), pp. 341-344.

TARABAN, M., ZHAN, H., WHITTEN, A.E., LANGLEY, D.B., MATTHEWS, K.S., SWINT-KRUSE, L. and TREWHELLA, J., 2008. Ligand-induced Conformational Changes and Conformational Dynamics in the Solution Structure of the Lactose Repressor Protein. *Journal of Molecular Biology,* **376**(2), pp. 466-481.

VELYVIS, A., SCHACHMAN, H.K. and KAY, L.E., 2009. Application of Methyl-TROSY NMR to Test Allosteric Models Describing Effects of Nucleotide Binding to Aspartate Transcarbamoylase. *Journal of Molecular Biology,* **387**(3), pp. 540-547.

WANG, X., VALLURUPALLI, P., VU, A., LEE, K., SUN, S., BAI, W., WU, C., ZHOU, H., SHEA, J., KAY, L.E. and DAHLQUIST, F.W., 2014. The Linker between the Dimerization and Catalytic Domains of the CheA Histidine Kinase Propagates Changes in Structure and Dynamics That Are Important for Enzymatic Activity. *Biochemistry,* **53**(5), pp. 855-861.

WELDON, J.E., RODGERS, M.E., LARKIN, C. and SCHLEIF, R.F., 2007. Structure and properties of a truely apo form of AraC dimerization domain. *Proteins,* **66**(3), pp. 646-654.

WILSON, C.J., ZHAN, H., SWINT-KRUSE, L. and MATTHEWS, K.S., 2007. Ligand interactions with lactose repressor protein and the repressor-operator complex: The effects of ionization and oligomerization on binding. *Biophysical chemistry; Julian Sturtevant Memorial Issue,* **126**(1), pp. 94-105.

WRIGGERS, W., CHAKRAVARTY, S. and JENNINGS, P.A., 2005. Control of protein functional dynamics by peptide linkers. *Biopolymers,* **80**(6), pp. 736-746.

YIFRACH, O. and MACKINNON, R., 2002. Energetics of Pore Opening in a Voltage-Gated K+ Channel. *Cell,* **111**(2), pp. 231-239.

YU, E.W. and KOSHLAND, D.E., J., 2001. Propagating conformational changes over long (and short) distances in proteins. *Proceedings of the National Academy of Sciences of the United States of America,* **98**(17), pp. 9517-9520.

YU, S., MAILLARD, R.A., GRIBENKO, A.V. and LEE, J.C., 2012. The N-terminal Capping Propensities of the D-helix Modulate the Allosteric Activation of the Escherichia coli cAMP Receptor Protein. *Journal of Biological Chemistry,* **287**(47), pp. 39402-39411.

ZHANG, X., REEDER, T. and SCHLEIF, R., 1996a. Transcription activation parameters at ara pBAD. *Journal of Molecular Biology,* **258**(1), pp. 14-24.

ZHANG, X., REEDER, T. and SCHLEIF, R., 1996. Transcription activation parameters at *ara* p_{BAD}. *Journal of Molecular Biology,* **258**(1), pp. 14-24.

Chapter 2: AraC orthologs reveal previously unknown intra-domain contacts

Abstract

In *E. coli* and many other bacteria, AraC acts as a transcription factor, either repressing or inducing the L-arabinose operon in the absence or presence, respectively, of its ligand, the plant sugar arabinose. Binding of arabinose causes a large-scale rearrangement of the protein, allowing different DNA binding patterns between the inducing and repressing state. Key to this action is a 20-residue N-terminal arm, which alters structure to bind over the sugar in the holo form of the protein. This arm is necessary for maintaining the repressive structure of the protein in the absence of arabinose, and the *E. coli* AraC arm tolerates very few mutations without disrupting repression. Despite this, arm sequences from orthologous AraC proteins vary significantly from the *E. coli* sequence. Chimeric AraC, comprising the *E. coli* sequence with the arm replaced by those of various orthologs, demonstrate a recovery of repression compared to the individual constituent point mutations. This indicates that the arm interacts primarily with itself during repression, rather than making functionally significant contacts elsewhere.

Introduction

The N-terminal arm of AraC – comprising the first 20 residues of the protein – is the region perhaps most altered in position between the apo and holo forms of the dimerization domain. In the absence of arabinose, the arm appears folded and positioned away from the sugar-binding pocket, closer to the region where the DNA-binding domain (DBD) is expected to lie(Weldon, Rodgers et al. 2007). In the presence of arabinose, the arm shifts to lie over the bound sugar(Soisson, MacDougall-Shackleton et al. 1997).

Since the arm undergoes a significant structural change between the inducing and repressing forms, and because of its proximity to the arabinose-binding pocket, it has long been an attractive candidate for attempts to identify the primary pathway of allosteric control within AraC. Indeed, the arm is clearly crucial for proper functioning of the protein. Removal of the arm results in constitutive expression from the *araBAD* genes, indicating a loss of repression. Furthermore, single-residue substitutions within the arm very frequently result in constitutive expression(Ross, Gryczynski et al. 2003), and very few residues outside this region display such a phenotype(Dirla, Chien et al. 2009).

Although it is clear that the arm plays a crucial role in repression, the mechanism by which it does so remains uncertain. Since it directly interacts with the sugar, the arm's repositioning is likely to be the initial structural consequence of ligand binding. Thus, the question becomes "where is the second link in the chain?" That is, what is the arm affecting or interacting with? As illustrated in Figure 1, any possible answer to this question must involve interaction with one of three possible locations: the DBD (Figure 1A), the dimerization domain (Figure 1B), or the linker (Figure 1C).

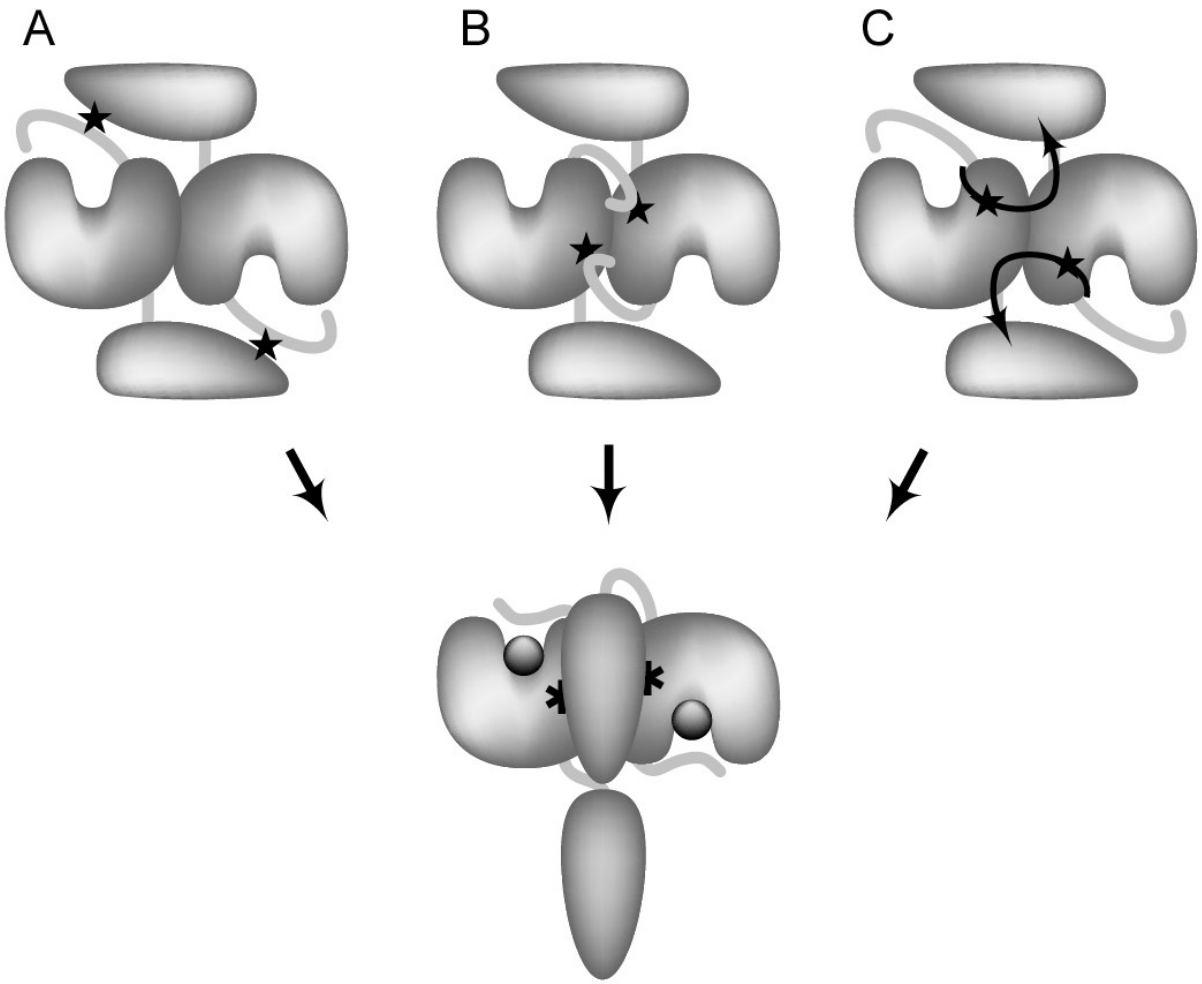

Figure 1. Possible mechanisms for arm control of repression. Star symbols indicate sites of potential interaction (A) Arm directly contacting DBDs, holding them in position. (B) Arm directly contacting dimerization domain, blocking contacts (bottom) with DBD. (C) Signal propagating through dimerization domain into linker of opposite subunit.

The first and perhaps the most initially attractive means by which the arm could be regulating the state of the protein is by direct interaction with the DBD (Figure 1A). The predicted structure of AraC in its repressing state places each DBD very near an N-terminal arm on the dimerization domain(Weldon, Rodgers et al. 2007, Rodgers, Schleif 2009b). This positioning allows for a specific interaction between the arms and DBDs that is controlled by the presence of arabinose. This interaction would "hold" the DBDs in positions facilitating repression, and breaking of these contacts – as would happen when the arm shifts to bind over

arabinose – would free the DBDs to relocate into orientations necessary for induction to occur. Such a mechanism would be consistent with the observed constitutive arm mutants(Dirla, Chien et al. 2009). On the other hand, one would predict analogous mutations on the DBD side of such contacts, but to date none such have been identified, nor have any DBD mutants been reported that rescue constitutive arm mutations(Ross, Gryczynski et al. 2003). Furthermore, several direct physical measurements have likewise failed to detect arm-DBD interactions(Rodgers, Holder et al. 2009).

A second potential mechanism would involve the arm interacting with the dimerization domain (Figure 1B). Rather than contacting the DBD, during repression the arm may instead be associating with the dimerization domain in such a way that it prevents the at least one DBD from moving into one of the positions required for induction. Since the AraC protein must contact adjacent DNA sites during induction(Lobell, Schleif 1990), and because of the limited length of the interdomain linker(Soisson, MacDougall-Shackleton et al. 1997), one DBD is constrained to lie over the dimerization domains in the inducing state, as shown in Figure 1. This conformation leaves these domains in such close proximity that interactions between the two are all but inevitable. This assumption is supported by previous work confirming that the dimerization domain does interact with the DBD during induction(Cole, Schleif 2012), though the exact locations have not been identified (this interdomain interaction will be further examined in Chapter 4). If the arm is interacting with the dimerization domain in such a way as to block these areas, it would act similar to a competitive inhibitor for the DBD, blocking the protein from assuming an inductive state until arabinose binding causes the arm to assume a new position.

The third possibility is that the arm acts via the linker (Figure 1C). This mechanism could involve either direct or indirect action. That is, either the arm residues could be directly interacting with those of the linker, or a conformational change in the arm could propagate a signal to the linker through adjacent regions of the protein (as suggested in Figure 1C). A direct interaction, however, is highly unlikely as the linker has been shown to be amenable to almost any substitution across its length without disrupting induction or repression(Malaga, Mayberry et al. 2016).

Previous work has shown that the relevant signal from the arm acts exclusively in *trans*, affecting the DBD of only the opposite subunit(Malaga, Mayberry et al. 2016). Crystal structures of the dimerization domain show that the arm is reasonably close to the opposite subunit's linker(Soisson, MacDougall-Shackleton et al. 1997, Weldon, Rodgers et al. 2007), allowing for multiple plausible pathways between these two regions. For example, local unfolding of the surface region between the base of the arm and the nearby linker could constitute a signal resulting in a structural change of the linker itself. Alternatively, the reorientation of the arm could propagate a series of structural changes through a region of the protein core into the linker. Such mechanisms are distinct from the previous two outlined above, as, during repression, they would not require direct functional contact between residues of the arm and those outside the arm.

To determine which of the above mechanisms are relevant to AraC repression, I replaced the N-terminal arm of *E. coli* AraC with the equivalent region from AraC proteins from other species. In addition to possessing sequence homology with *E. coli* AraC, these orthologs are each expressed from genes located within the L-arabinose operon of their respective bacteria's genomes, signifying their conserved roles as transcriptional regulators. Thus, this work and its conclusions are based on the assumption that the general mechanisms employed by these AraC orthologs are conserved relative to *E. coli* AraC.

In all but one case, the new arms contained mutations previously shown to yield constitutive expression in the context of *E. coli* AraC(Ross, Gryczynski et al. 2003) (Figure 2). Presumably, in the variants chosen for this study, secondary variations must compensate for variations that cause constitutivity, assuming these arms work in their native context. The question I am addressing in this work is that of the location of these secondary mutations. If they lie in the arm itself, it would suggest that the arms do not make critical, functionally important contacts elsewhere in the protein. Conversely, if the compensating mutations are located in other regions, it would strongly suggest that such functionally significant interdomain contacts occur. Thus, if the chosen variant arms, when substituted into *E. coli* AraC, repress, then these compensating mutations being sought must lie within the arm itself.

Results

In vivo behavior of AraC orthologs

To probe for the existence of intra-arm contacts affecting repression, I used a series of chimeric AraC proteins wherein I replaced the first 19 residues of the *E. coli* AraC sequence with the corresponding residues of orthologous proteins from various bacterial genera. Orthologs were chosen based on their inclusion of single-residue polymorphisms within the arms that had been previously identified individually as exhibiting a loss of repression in *E. coli* AraC(Dirla, Chien et al. 2009). Assuming that these arms function properly in their original host, these mutations must be compensated by other mutations somewhere in the protein. If the orthologous arms function properly substituted into in *E. coli* AraC, then the compensating mutations must reside within the arm itself, indicating that an intra-arm interaction is necessary for repression. If the arm is making inter-domain contacts, however, then the compensating mutations exist elsewhere in the protein, and the substituted arm would not result in normal repression, as the necessary alterations to the contacted site are unlikely to be present in AraC.

Figure 2 shows the results of testing six different arms. The first, originating from *Cronobacter*, contains no mutations known to be constitutive in *E. coli*. This sequence does, however, contain a considerable N-terminal extension. The *Cronobacter* arm was examined as a control, in order to confirm that such an extension confers no phenotypic alteration to *E. coli* AraC, a necessary point because several of the other arms studied include similar extensions.

In 4 out of 5 cases, I observed the recovery of significant levels of repression compared to the individual point mutants (Figure 2). In 3 of these cases (*Erwinia*, *Aeromonas*, and *Leptothrix*), arm hybrids induced and repressed about like wild type AraC, despite containing mutations known to be constitutive in the fully *E. coli* context(Dirla, Chien et al. 2009). The *Erwinia* and *Leptothrix* arm sequences are of particular note, as they each contain only a single mutation that causes the *E. coli* AraC to become to be constitutive. All three of these sequences have N-terminal extensions as well as mutations in the first 5 residues from the *E. coli* sequence, strongly suggesting an interaction with this early region. This region

was previously shown to be disposable in *E. coli* AraC, with a deletion of the first 5 residues showing normal induction(Eustance, Bustos et al. 1994).

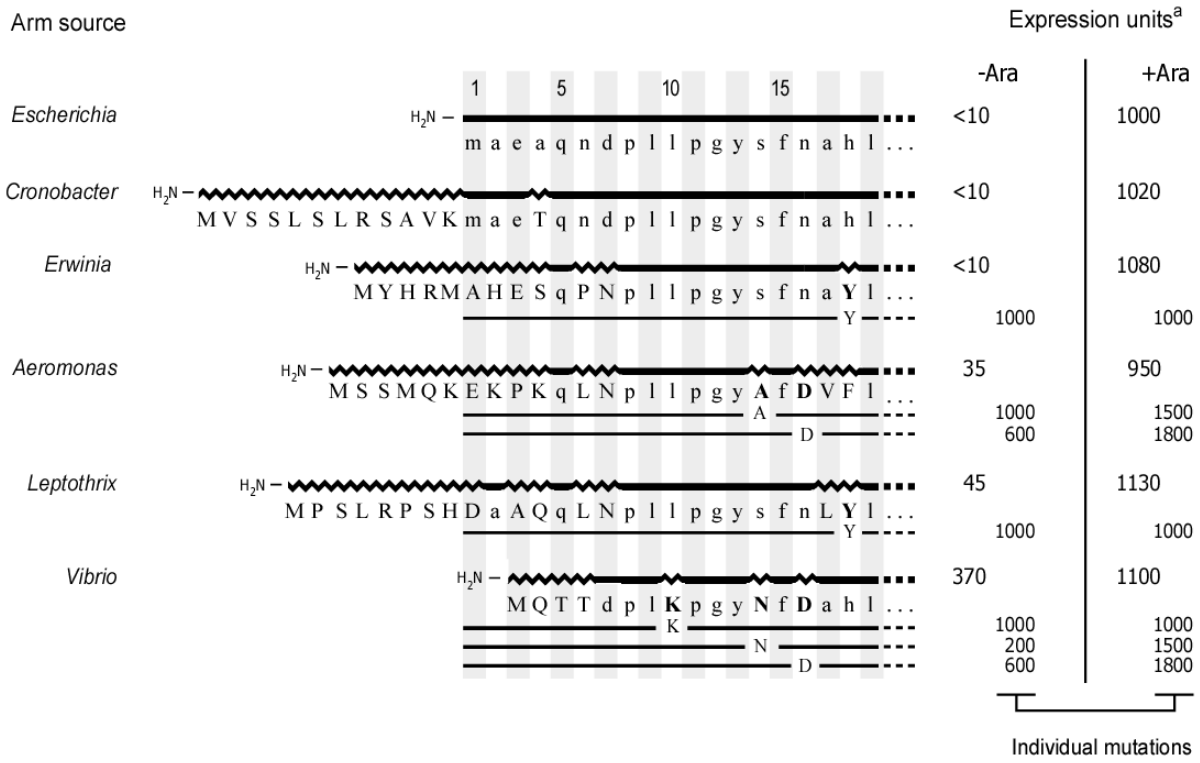

Figure 2: *In vivo* expression of arm mutants in *E. coli* AraC. All sequences measured in context of *E. coli* AraC. Straight lines/lowercase letters represent residues identical to the *E. coli* AraC protein sequence; jagged lines/capital letters indicate altered residues. Expression levels of single-residue point mutants measured by Ross et al(Ross, Gryczynski et al. 2003). Bold letters represent mutations known to result in constitutive expression in *E. coli*.

[a]Values are the average of multiple independent measurements of arabinose isomerase levels in AraC$^-$ cells expressing the indicated form of *araC* from a plasmid.

Discussion

In the work presented here, I have shown that, within the N-terminal arm of AraC, constitutive mutations can be rescued solely by further changes within the arm. Replacement of the N-terminal arm of AraC with an orthologous arm preserved the protein's ability to repress in the absence of arabinose. This occurred despite the presence of mutations of residues within the *E. coli* arm that would otherwise eliminate

the protein's ability to properly repress. Since the arm is the only region altered in these hybrid proteins, this demonstrates that, at least in the case of these residues, the function vital for repression is the overall structure of the arm, maintained by establishment of intra-arm contacts.

The ortholog arm substitution experiments strongly suggest that the *E. coli* N-terminal arm of AraC does not normally make contacts with the DBD that are required for repression. This conclusion is in accord with a number of previous experimental results. First, although a change in virtually any residue of the arm generates constitutivity(Ross, Gryczynski et al. 2003), no mutations with similar effects have been identified in the DBD(Dirla, Chien et al. 2009). Second, direct physical measurements, plasmon resonance, and NMR(Rodgers, Holder et al. 2009) have likewise failed to detect any arm-DBD interactions. In light of such plentiful evidence, it is highly unlikely that the arm makes any functionally important interactions with the DBD.

The work presented here likewise suggests that the arms do not make meaningful interactions with the dimerization domains for repression. For this possibility, however, there is less auxiliary experimental support. While plasmon resonance experiments have detected arm-dimerization domain interactions in the presence of arabinose, they detected none in the sugar's absence(Ghosh, Schleif 2001). Constitutive mutations have been found in the dimerization domain(Dirla, Chien et al. 2009). NMR experiments analogous to those performed on the DBD have not been carried out, as the dimerization domain is large, and poorly soluble in the absence of arabinose, making such experiments prohibitively costly and difficult. Thus, while we consider functional arm-dimerization domain interactions unlikely in the repressing state, we cannot conclusively rule out the possibility.

The results presented here are consistent with a mechanism involving the propagation of a structural signal through the protein to linker. Previous experiments have shown that the linker – even without the DBD attached – alters structure in response to arabinose binding, and the relevant allosteric signal is communicated in *trans,* to the linker of the opposite subunit(Malaga, Mayberry et al. 2016). A direct interaction between the arm and the linker is unlikely, however, as the linker itself is able to tolerate almost

any substitution(Malaga, Mayberry et al. 2016, Seedorff, Schleif 2011). Thus, a likely mechanism involves a structured arm in the absence of arabinose crowding nearby residues in the dimerization domain in such a way as to maintain a particular linker structure; when arabinose binds, the restructuring of the arm over it would allow a rearrangement of these adjacent residues, proliferating along a path towards the adjacent linker, culminating in a structural change therein. This change would then cause or allow the DBDs to reorient into the inducing state of the protein. The role of the linker's structure in this mechanism is more deeply explored in Chapter 3.

While three of the arm mutants showed WT or nearly-WT repression levels, the chimera including the arm from *Vibrio* still showed diminished repression. In the case of this arm, which contains the L10K and N16D variations, it should be noted that this phenotype still represents a partial rescue from that of the individual point mutants. AraC with the L10K mutation displays 100% expression in both the presence and absence of arabinose, while the N16D mutation shows 60% expression in the absence of arabinose (Figure 2). Thus, the observed 37% expression in the *Vibrio-Escherichia* chimeric AraC without arabinose represents a partial recovery of repression. These results may indicate that arm structure does not represent the whole story where repression is concerned, and there may be contacts outside this region, though it is also possible that the *Vibrio* AraC protein has evolved a subtly different mechanism, and such contacts are not relevant in *E. coli*, serving only to disrupt the necessary intra-arm interactions in the chimera.

The results presented here, in concert with previous findings, lead to the rejection of the notion of arm binding to either the DBD or dimerization domain during repression in AraC. Thus, it is highly likely that the relevant mechanism involves the arm signaling through the linker. The manner of this control via the linker will be explored further in the following chapter.

Materials and Methods

Plasmids and strains

AraC arm mutants were generated using Stratagene QuikChange™ protocol for site-directed mutagenesis (Stratagene, La Jolla, CA) from oligonucleotide primers and verified by sequencing the products. In all arm mutants except *Cronobacter*, mutagenesis reactions were carried out stepwise in 2 phases, mutagenizing one portion first, and using the product of this reaction as the template for the second reaction. The largest sequence inserted in one reaction was 34nt.

For the *in vivo* activity measurements, AraC the protein was expressed from the variable copy number plasmid pCCN AraC expression vector based on the p_{BAD}-GFP plasmid. This variable copy number plasmid contains two origins of replication: f1 which will maintain a single copy per cell and oriS which allows for high copy-number but is blocked by LacI if present at elevated levels. For measurement of activity from AraC variants carried on this plasmid, it was transformed into a strain deleted of AraC, but containing *ara* p_{BAD} and the *araBAD* genes, thereby allowing measurement of arabinose isomerase for quantitation of the p_{BAD} basal levels. This strain, RS321, is derived from SH321 (*F⁻ ΔaraC-leu1022 Δlac74 galK⁻ Str^r thi⁻1⁴*)(Hahn, Dunn et al. 1984) by P1 transduction transferring Tn10(laci^{Q} Cm^{r} at the *lac* locus) from Gottesman strain ASP7020 (NM514)(Parker, Gottesman 2016).

Arabinose isomerase assay

Arabinose isomerase assays were performed as described by Schleif and Wensink(Schleif, Wensink 1981). Cells, containing the plasmid, SH321 (*F⁻ ΔaraC-leu1022 Δlac74 galK⁻ Str^r thi⁻1⁴*) or SH322 (*F⁻ Δlac74 galK⁻ Str^r thi⁻1⁴*)(Hahn, Dunn et al. 1984) were grown in M10 minimal salts medium(Schleif, Wensink 1981) plus 0.4% glycerol or 0.4% arabinose, 10 µg/ml thiamine, 0.2% casamino acids, and 100 µg/mL ampicillin, to an apparent OD_{650} of 0.6 to 0.8 (as measured by a Perkin/Elmer Lambda 25 UV/Vis spectrophotometer) before harvesting and concentration for assay of isomerase. Cells containing pCCNC were grown to stationary phase in YT medium containing 40 µg/ml kanamycin for measurement of uninduced levels of arabinose isomerase.

References

COLE, S.D. and SCHLEIF, R., 2012. A new and unexpected domain-domain interaction in the AraC protein. *Proteins: Structure, Function, and Bioinformatics,* **80**(5), pp. 1465-1475.

DIRLA, S., CHIEN, J.Y. and SCHLEIF, R., 2009. Constitutive mutations in the *Escherichia coli* AraC protein. *Journal of Bacteriology,* **191**(8), pp. 2668-2674.

EUSTANCE, R.J., BUSTOS, S.A. and SCHLEIF, R.F., 1994. Reaching out. Locating and lengthening the interdomain linker in AraC protein. *Journal of Molecular Biology,* **242**(4), pp. 330-338.

GHOSH, M. and SCHLEIF, R.F., 2001. Biophysical evidence of arm-domain interactions in AraC. *Analytical Biochemistry,* **295**(1), pp. 107-112.

HAHN, S., DUNN, T. and SCHLEIF, R., 1984. Upstream repression and CRP stimulation of the *Escherichia coli* L-arabinose operon. *Journal of Molecular Biology,* **180**(1), pp. 61-72.

LOBELL, R.B. and SCHLEIF, R.F., 1990. DNA looping and unlooping by AraC protein. *Science (New York, N.Y.),* **250**(4980), pp. 528-532.

MALAGA, F., MAYBERRY, O., PARK, D., RODGERS, M.E., TOPTYGIN, D. and SCHLEIF, R.F., 2016. A genetic and physical study of the interdomain linker of E. Coli AraC protein-a trans-subunit communication pathway. *Proteins: Structure, Function, and Bioinformatics,* **84**, pp. 448--460.

PARKER, A. and GOTTESMAN, S., 2016. Small RNA Regulation of TolC, the Outer Membrane Component of Bacterial Multidrug Transporters. *J. Bacteriol.,* **198**(7), pp. 1101-1113.

RODGERS, M.E., HOLDER, N.D., DIRLA, S. and SCHLEIF, R., 2009. Functional modes of the regulatory arm of AraC. *Proteins,* **74**(1), pp. 81-91.

RODGERS, M.E. and SCHLEIF, R., 2009. Solution structure of the DNA binding domain of AraC protein. *Proteins,* **77**(1), pp. 202-208.

ROSS, J.J., GRYCZYNSKI, U. and SCHLEIF, R., 2003. Mutational analysis of residue roles in AraC function. *Journal of Molecular Biology,* **328**(1), pp. 85-93.

SCHLEIF, R.F. and WENSINK, P.C., 1981. *Practical methods in molecular biology.* New York: Springer-Verlag.

SEEDORFF, J. and SCHLEIF, R., 2011. Active role of the interdomain linker of AraC. *Journal of Bacteriology,* **193**(20), pp. 5737-5746.

SOISSON, S.M., MACDOUGALL-SHACKLETON, B., SCHLEIF, R. and WOLBERGER, C., 1997. The 1.6 A crystal structure of the AraC sugar-binding and dimerization domain complexed with D-fucose. *Journal of Molecular Biology,* **273**(1), pp. 226-237.

WELDON, J.E., RODGERS, M.E., LARKIN, C. and SCHLEIF, R.F., 2007. Structure and properties of a truely apo form of AraC dimerization domain. *Proteins,* **66**(3), pp. 646-654.

Chapter 3: Helicity in the Interdomain Linker in AraC

Abstract

In *Escherichia coli*, the dimeric AraC protein actively represses transcription from the L-arabinose *araBAD* operon in the absence of arabinose, but induces transcription in its presence. Here I provide evidence that, in shifting from the repressing to the inducing state, the interdomain linker shifts from an alpha-helix to a flexible coil. *In vivo* and *in vitro* experiments show that AraC with a linker sequence that favors helix formation is shifted towards the repressing state both in the absence and presence of arabinose. Conversely, AraC containing a linker sequence that is unfavorable for helix formation is shifted towards the inducing state. Experiments in which the presumed helical linker is shortened or lengthened – protein helical-twist experiments – also support the helix-coil mechanism. Previous experiments have shown that, upon the binding of arabinose, the apparent rigidity with which the DNA binding domains of AraC are held in space decreases. Thus, arabinose likely controls the stability of the interdomain linker's helicity. Circular dichroism experiments with peptides show that helicity of the linker sequence can be controlled by the helicity of residues preceding the linker, providing a plausible mechanism for arabinose to control the repressing-inducing state of AraC protein.

Introduction

Interdomain linkers have been studied in many proteins and have been implicated in numerous roles, including tethering, and signal propagation(Wang, Vallurupalli et al. 2014, Ma, Tsai et al. 2011, Kukic, Camilloni et al. 2014, Wriggers, Chakravarty et al. 2005). The AraC protein presents a case where the linker structure is directly involved with a large-scale structural change that dramatically affects the DNA binding activity of the protein.

In *E. coli*, expression of the genes required for the catabolism of the sugar L-arabinose, *araBAD*, is controlled by the *araC* gene product(Sheppard, Englesberg 1967). AraC functions as a homodimer, with each subunit composed of two major domains(Bustos, Schleif 1993) (Figure 1). The first is a 166-residue

dimerization domain, the first 20 residues of which comprise an N-terminal arm. The C-terminal domain is a 117 residue DNA-binding domain. This DBD is connected to the dimerization domain by an 8-residue interdomain linker(Soisson, MacDougall-Shackleton et al. 1997, Rodgers, Schleif 2009).

In the absence of arabinose, the AraC dimer contacts the I_1 DNA half-site and the O_2 half-site lying 210 base pairs upstream from I_1(Dunn, Hahn et al. 1984) (Figure 1). AraC binding to I_1 and O_2 loops the DNA. This interferes with the binding of RNA polymerase to the nearby *ara* p_{BAD} promoter(Hahn, Schleif 1983), thereby repressing its activity. A third binding site, I_2, lies adjacent to I_1 in a direct repeat orientation, partially overlapping the p_{BAD} core promoter (Carra, Schleif 1993, Dunn, Hahn et al. 1984). It is unoccupied in the minus-arabinose, repressive, state of the operon(Martin, Huo et al. 1986), but when arabinose is added, the DNA binding domain previously at the O_2 half-site relocates to the I_2 half-site, eliminating the DNA looping(Lobell, Schleif 1990). Following its repositioning on the DNA, AraC stimulates *araBAD* expression from the p_{BAD} promoter(Hendrickson, Schleif 1985, Lobell, Schleif 1990) by assisting the DNA binding and isomerization of RNA polymerase to form an open complex(Zhang, Reeder et al. 1996). Therefore, DNA looping hinders activity of p_{BAD} in two ways: first, by interfering with polymerase access to the promoter, and second, by holding a DNA binding domain of AraC from binding to I_2.

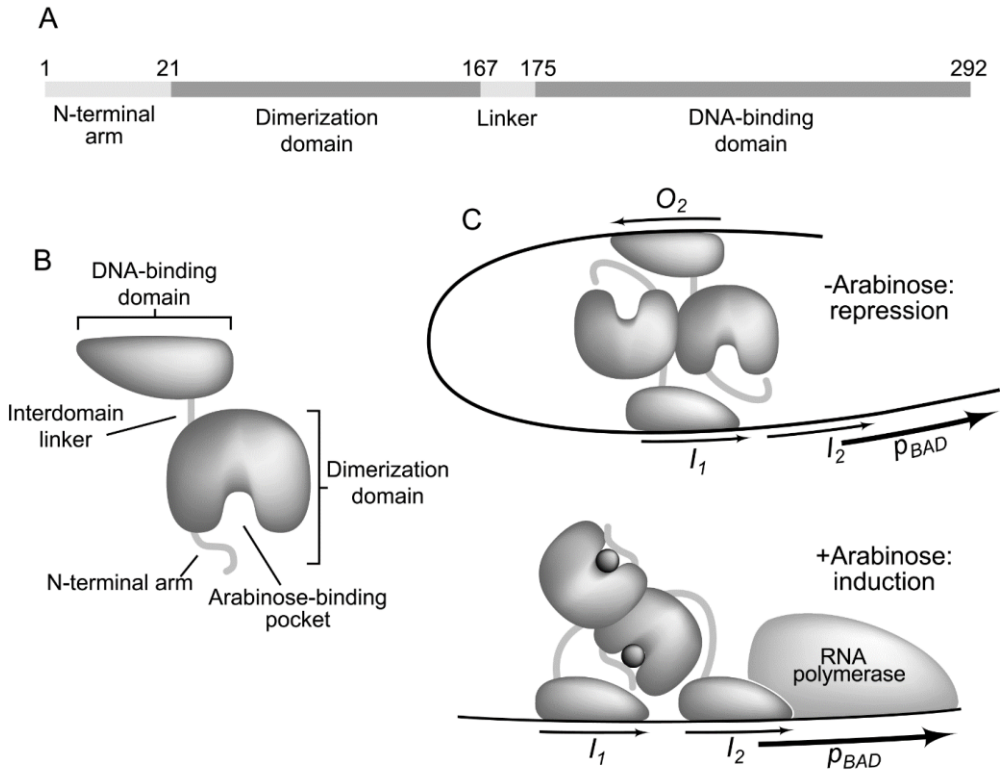

Figure 1. (A) Linear layout of the AraC protein domains (B) Domain structure of AraC. (C) The DNA binding sites in the *ara* regulatory region made by AraC in the absence and presence of arabinose. The p_{BAD} is the RNA polymerase binding site

Previous experiments have shown that reversing the direction of the asymmetric O_2 half-site while maintaining its position on the same face of the DNA reduces or eliminates the capacity of AraC to loop to O_2 from I_1(Seabold, Schleif 1998). This plus *in vitro* DNA binding experiments using direct repeat and inverted repeats of the half-sites(Carra, Schleif 1993) show that in the absence of arabinose, the relative positions and/or orientations of the DNA binding domains of AraC are constrained in a way that favors DNA looping between O_2 and I_1 as well as hindering their binding to direct repeat half-sites. On the binding of arabinose, this apparent rigidity is reduced, thus changing the lower energy state of the system from the repressive, DNA looping, state to the inducing state with AraC bound to the I_1-I_2 direct repeat DNA half-sites.

The abovementioned facts raise two fundamental mechanistic questions: how is the rigidity that maintains the orientations of the DNA binding domains in the absence of arabinose generated, and then, how is the rigidity relaxed when arabinose is added? The N-terminal 20 amino acids of AraC are an attractive candidate for controlling the rigidity. The arms' locations in the protein's three-dimensional structure are compatible with their binding to the DBDs in the absence of arabinose, and this interaction could then hold the DNA binding domains in positions suitable for DNA looping(Soisson, MacDougall-Shackleton et al. 1997). Furthermore, in the absence of arabinose, the arms are available to bind to the DNA binding domains (Weldon, Rodgers et al. 2007), but in the presence of arabinose, the arms are draped over the dimerization domain, with the side chains of F15 reaching into the domains and making direct contact with the bound arabinose(Soisson, MacDougall-Shackleton et al. 1997).

Although arm-DNA binding domain interactions in the absence of arabinose appear to be an attractive regulatory mechanism, multiple experiments to detect such contacts have failed to do so (to be discussed below). Experimental attention has therefore shifted from mechanisms involving direct arm-DBD interactions to less direct mechanisms for holding the DNA binding domains. *In vitro* experiments have recently suggested that the interdomain linker plays a direct role in the rigidity shift of AraC(Malaga, Mayberry et al. 2016). These experiments utilized truncated AraC consisting of dimerization domain, including the N-terminal arm and the eight-residue interdomain linker, but lacking the DNA binding domain. Fluorescence anisotropy measurements using such a truncated AraC with a fluorescent label at the C-terminal end of the linker showed that the label's rotational motion is controlled by arabinose(Malaga, Mayberry et al. 2016). In the presence of arabinose, its motion is increased, as though the linker's rigidity is reduced by the presence of arabinose. Additionally, a linker's apparent flexibility is controlled by only that N-terminal arm that is adjacent to the linker, and not by the other arm, implying that it is a relatively direct interaction between arm and linker that controls flexibility rather than a global structural change.

In the work described below, I address the genetic hint contained in earlier work suggesting that, in the repressing state, the interdomain linker is largely in an alpha-helical form, and the presence of arabinose

shifts the linker towards a random coil state(Seedorff, Schleif 2011). This mechanism is well suited to the rigidity shift mentioned above. Here I show that increasing the helical propensity of the linker increases the tendency of AraC to be in the repressing state as shown by a reduction in the inducibility by AraC *in vivo* and by a reduction in the affinity for binding to adjacent direct repeat half-sites (I_1-I_2) *in vitro*. Conversely, decreasing the helical propensity of the linker decreases the tendency to be in the repressing state as shown by increases in *in vivo* inducibility and uninduced basal levels, as well as increasing the protein's *in vitro* DNA binding affinity to direct repeat half-sites. I have also obtained circular dichroism spectra showing that, although a peptide reproducing the linker sequence is largely unstructured in solution, it becomes helical when a longer helical peptide sequence is added to it. In AraC, the linker immediately follows a helix in the protein's dimerization domain, and hence, the dimerizing helix could nucleate helix formation of the linker in the protein. Furthermore, this suggests that control of the linker's structural state may be modulated by that of the immediately preceding residues.

Results

Helix stabilizing and destabilizing mutants – *in vivo* properties

To increase the helical preference of the eight-residue linker (WT sequence INESLHPP), I replaced its first six residues with alanine, the residue with the highest individual helical propensity(Pace, Scholtz 1998), giving a linker sequence of AAAAAAPP (6A). Conversely, I also constructed a glycine-serine linker, GGGGSGPP (GS), to decrease the helical propensity of the sequence and bias its structure towards a random coil(Pace, Scholtz 1998).

Table 1 shows the repressing and inducing activities of wild type AraC and AraC containing the 6A and GS linkers. In the case of AraC with the 6A, helix-favoring linker, uninduced levels remained comparable to those seen in WT AraC, while induced levels were reduced by a factor of two. Conversely, the GS, random coil-favoring, linker mutant showed both a greatly elevated uninduced level – that is a loss

of repression – and an elevated induced level. These results are consistent with the idea that the linker is helical in the repressing state and is not helical in the inducing state.

Table 1 *In vivo* activity of AraC linker mutants on $arap_{BAD}$[a] in units per cell(Schleif, Wensink 1981)

	-arabinose	+arabinose
WT	4.6 ± 1.7	960 ± 190
6A	5.4 ± 1.6	510 ± 110
GS	130 ± 5	3100 ± 230

[a]Values are the average and standard deviation of three independent measurements of arabinose isomerase levels in $AraC^-$ cells expressing the indicated form of *araC* from a plasmid.

Excluding a potential artifact

The reduced ability of AraC containing the 6A linker to induce could result from one of three possibilities: (1) as in my hypothesis, a shift towards the repressing conformation in the arabinose-controlled equilibrium between the repressing and inducing conformations, (2) reduced total levels of active protein, (3) a reduced ability to activate transcription once bound at *araI*.

It has previously been shown that the repressing state of AraC is dominant to the inducing state(Sheppard, Englesberg 1967), which is an expected consequence of the DNA looping mechanism of the system. Therefore, if the 6A AraC has been pushed towards the repressing state, its hypo-inducible phenotype will be dominant over the fully inducible wild type. Such dominance would not be seen for the other two possible reasons for hypo-inducibility. Table 2 shows that hypo-inducibility is dominant, and hence the behavior of the linker variants is consistent with the helix-coil hypothesis.

Table 2 Hypoinducibility of the 6A linker is dominant to wild type AraC[a].

AraC linker	-arabinose	+arabinose
WT	4.3 ± 0.9	1100 ± 235
6A	6.0 ± 2.2	625 ± 80

[a]Units arabinose isomerase per cell from three independent measurements in heterozygous cells containing wild type *araCBAD* genes on the chromosome and the indicated mutant AraC on the plasmid.

In vitro properties of the helix stabilizing and destabilizing variants

As mentioned earlier, the binding affinity of AraC *in vitro* to adjacent direct as well as inverted half sites has been studied(Carra, Schleif 1993). The results indicate that the AraC DNA binding domains are significantly constrained in the absence of arabinose to positions and orientations that are unfavorable for binding to adjacent, DNA half-sites, but that in the presence of arabinose, they are less constrained. Thus, in the absence of arabinose, the net free energy available for binding to adjacent DNA half-sites is reduced by the amount of energy required to overcome the constraints. In the presence of arabinose, less energy is required to overcome the constraints, and AraC can bind the adjacent direct repeat DNA more tightly. Consequently, if the interdomain linker in wild type AraC in the repressing state were helical and breaks in order for the protein to bind to the DNA sites required for induction, then an interdomain linker sequence that increases the stability of the helix will weaken the binding of AraC to adjacent DNA half-sites and a linker sequence that decreases the stability of the helix will do the opposite.

For AraC and many other DNA binding proteins whose forward rate of DNA binding is constant and close to the diffusion limit(Berg, von Hippel 1985, Gutfreund 1995, Hendrickson, Schleif 1984), relative equilibrium binding constants can be determined by comparing their more easily measured dissociation

rates. I used a direct repeat of the I_1 half-site to further facilitate measurements as AraC dissociates more slowly from this than from the natural *I* site, (I_1-I_2)

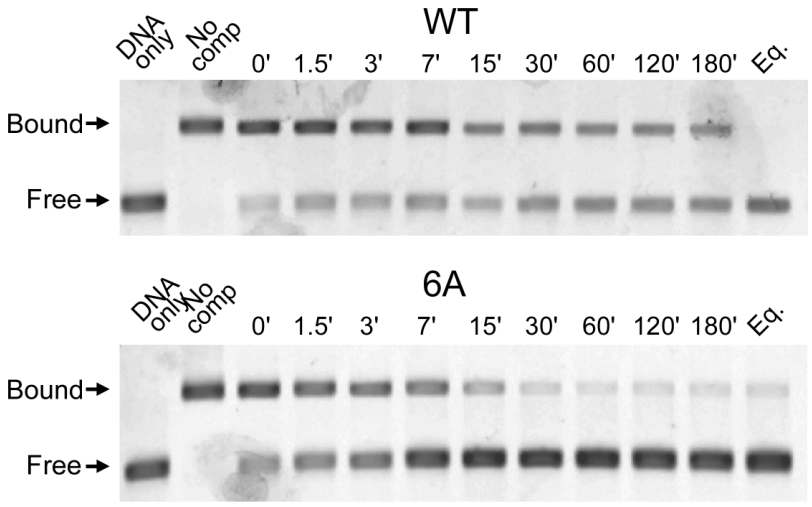

Figure 2. DNA electrophoretic mobility shift measuring the dissociation kinetics of AraC protein from I_1-I_1 site in the presence of arabinose and 300 mM KCl. First lane contains only labeled I_1-I_1 DNA; second lane contains labeled I_1-I_1 DNA and protein to show that DNA is completely bound; for lanes 3-11 samples were incubated with unlabeled I_1-I_1 for the indicated time before being loaded into the gel; for the sample run in the final lane, competitor DNA was added before the protein to verify that, in the other samples, sufficient competitor was present to assure that any protein that dissociated from labelled DNA would almost surely reassociate with the competitor and not the labelled DNA.

Figure 2 shows a representative gel from which dissociation kinetics can be determined, and Figure 3 shows the dissociation kinetics of wild type and the 6A and GS mutant linker proteins in the absence and presence of arabinose. In the absence of arabinose, the helix-favoring 6A AraC dissociates from DNA significantly faster than wild type AraC. In the presence of arabinose, the 6A AraC dissociates at about the same rate as wild type AraC. Conversely, the random coil favoring GS AraC dissociates more slowly from DNA than wild type AraC, both in the absence and presence of arabinose. These results, also, are consistent with the idea that a linker sequence favoring a helical conformation weakens binding to adjacent DNA half-sites, and are consistent with the *in vivo* results. That is, in the presence of arabinose, the equilibrium state of the linker shifts toward a random coil, which facilitates binding to adjacent half-sites.

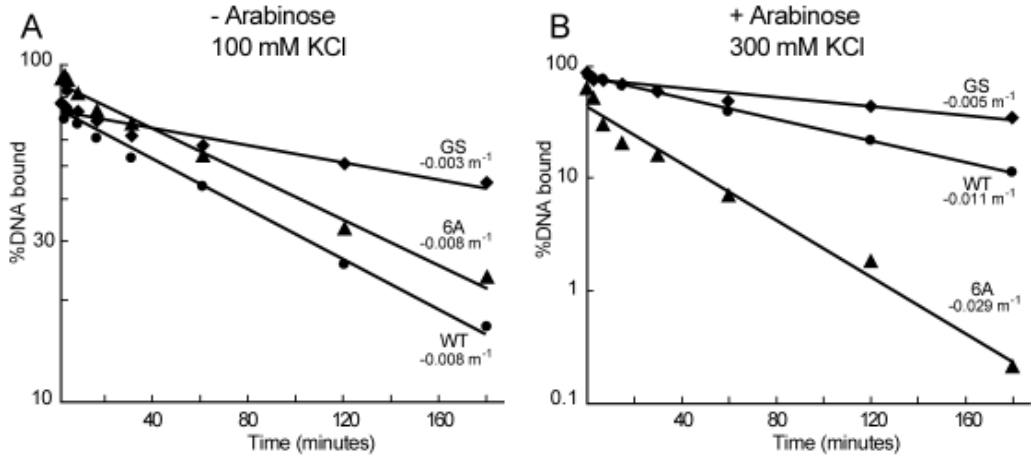

Figure 3. Dissociation of I_1-I_1 from AraC linker variants as a function of time. **A.** Dissociation in 100 mM KCl in the absence of arabinose. **B.** Dissociation in 300 mM KCl in the presence of arabinose. Plus- and minus- arabinose measurements were done at different salt concentrations so as to maintain experimentally convenient dissociation rates.

Helix propagation into the linker sequence

A sequence of six or eight residues is unlikely to be able to nucleate and form a helix on its own. If, however, it were directly preceded by a helix, as is the case in AraC, it could well be helical itself. I therefore sought to determine if a preceding helix can be extended into the linker, as would be required for my helix-coil transition model. To test this question, I used a series of peptides and examined their helical content by CD. This also allowed me to test a corollary: whether or not the linker, when connected to the C-terminus of an existing helix, is itself helical by default, or if it would require an external signal to further stabilize the helix.

I used four peptide sequences: the eight-residue linker alone, a 16-residue helical sequence alone, and this helix attached to the linker by either an alanine or a proline residue (Table 3). I expected the helix-Ala-linker sequence to allow the helix structure to continue into the linker, and the helix-Pro-linker to interrupt it. My pre-nucleated "seed" helix was based on the 3K(I) sequence characterized by Marqusee *et al*(Marqusee, Robbins et al. 1989), differing only by my inclusion of an N-terminal tryptophan residue for spectrophotometric concentration measurement. The CD spectra of these peptides are shown in Figure 4. The linker sequence alone displayed a character consistent with being almost entirely unstructured. Conversely, the helix sequence showed a highly helical spectrum, as previously reported(Marqusee, Robbins et al. 1989). The helix-Ala-linker sequence showed a spectrum almost perfectly overlapping that of the helix alone, that is, no evidence of any additional random coil structure. Note particularly the two local minima around 222 nm and 208 nm, as well as in the maximum around 190(Brahms, Brahms 1980). I therefore conclude that most of the linker sequence in this peptide is also helical. Notably, a sizeable shift in

Table 3 Sequence layout of linker-capped peptides

	Helix	Linker
Linker	Ac-	INESLHPPW
Helix	Ac-WAAAAKAAAAKAAAAKA	
H-A-L	Ac-WAAAAKAAAAKAAAAKA**A**INESLHPPW	
H-P-L	Ac-WAAAAKAAAAKAAAAKA**P**INESLHPPW	

Ac – acetyl group. Boldface indicates residue not included in either linker or helix sequence.

the CD spectrum is seen when the linker sequence is preceded by a proline rather than an alanine, indicating that the introduction of this proline disrupts helix formation either in the preceding or following helix.

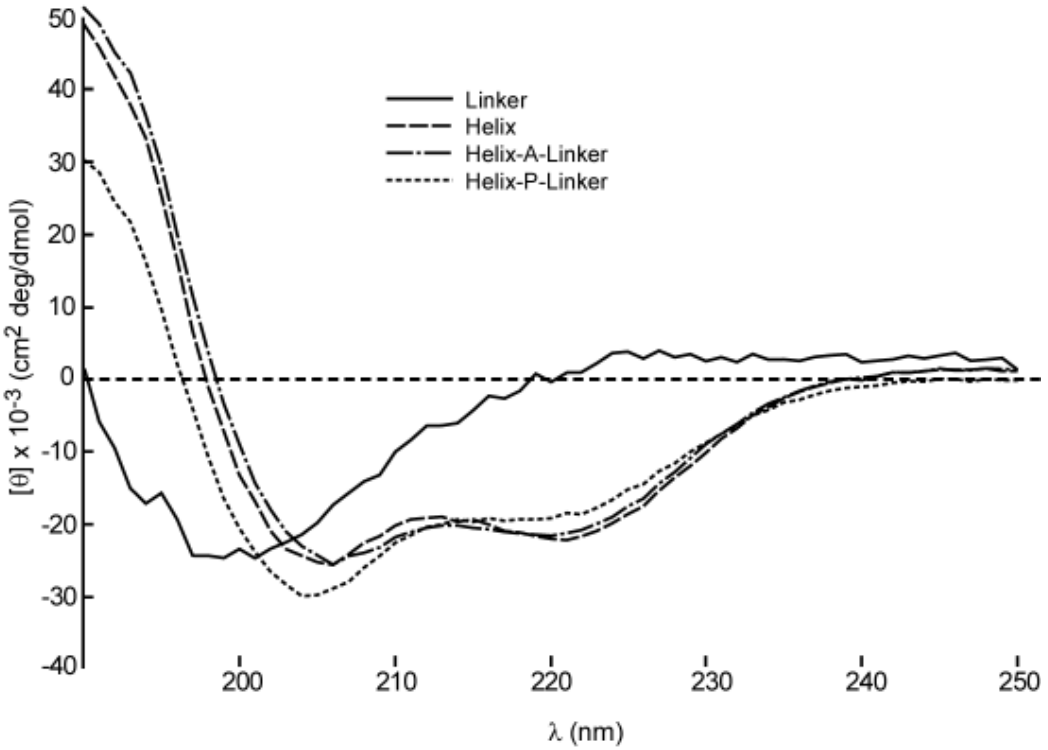

Figure 4. CD spectra of helix/linker peptides. All measurements were done in 10 mM phosphate buffer pH 7, 0.1 M NaCl at 2.0° C

Discussion

I have investigated what happens when AraC protein shifts from its repressing and DNA looping state in the absence of arabinose to its nonlooping and inducing state in the presence of arabinose(Lobell, Schleif 1990, Martin, Huo et al. 1986). In the apo state, the DNA binding domains of AraC apparently are rigidly held in positions and orientations that favor binding to two well-separated DNA half-sites over binding to two adjacent DNA half-sites. The addition of arabinose reverses the binding site preference of AraC by reducing the energetic cost of binding to adjacent DNA half-sites. Thus, the protein shifts from repressing the *ara* p_{BAD} promoter by DNA looping to inducing it by binding to the adjacent I_1 and I_2 half-

sites(Lobell, Schleif 1990, Martin, Huo et al. 1986). This chapter provides several types of experiments indicating that in the absence of arabinose, the interdomain linkers between the dimerization domains and DNA binding domains are largely in an alpha helical state, and that, upon binding arabinose, the linker is not in a helical state. These findings, in conjunction with those found previously(Malaga, Mayberry et al. 2016), suggest a mechanism whereby the induction status of AraC is controlled by a helix-coil transition in the interdomain linker.

While direct examination of the structure of the linker would be ideal, a number of factors hinder the feasibility of such investigation. Structure determination on full length AraC by crystallography is not available as more than 30 years of serious efforts to crystallize the protein have not been successful. Structure determination of the 64,000-molecular weight AraC protein by NMR, while possible would be highly expensive, and most challenging due in part to the low solubility of apo AraC. While I do present CD spectra using peptides reproducing the linker sequence here, using similar means to monitor the helical status of the linker in full length protein is unfeasible because 37% of the 292 residue AraC protomer is already helical, thereby preventing detection of such small changes.

To test whether the interdomain linker is helical in the repressing state and nonhelical in the inducing state, I both increased and decreased the stability of an alpha-helical state of the linker. I investigated the *in vivo* regulatory properties and *in vitro* DNA binding properties of wild type AraC and of two variants: one whose interdomain linker favors helix formation, and one whose linker disfavors helix formation. Compared to wild type AraC whose linker sequence is INESLHPP, the protein containing a helix-stabilizing(Pace, Scholtz 1998) alanine-substituted linker sequence, AAAAAAPP, shifts AraC toward its repressing state and weakens its binding to adjacent direct repeat half-sites. AraC containing a sequence disfavoring helix formation(Pace, Scholtz 1998), GGGGSGPP, produced the opposite effect. These results are consistent with the idea that the shift from repression to induction involves a transition of at least a portion of AraC's interdomain linker from an alpha-helix to a flexible coil, and the fraction of protein in either state can be increased by appropriate substitutions in the linker.

My spectroscopic examination of peptides shows that the linker sequence itself, when free in solution, and thus absent any other interactions, does not show any significant helical content. However, when added to the end of a helix, it adopts a helical character emulating that of the nucleating helix itself. Thus, it is clear that at least most of the linker is capable of adopting a helical conformation when propagated from the C-terminus of such a structure. Indeed, this is precisely the environment found in the AraC protein itself, with the linker directly following the second helix of a coiled-coil in the dimerization domain(Soisson, MacDougall-Shackleton et al. 1997). This provides a potential method of control of linker helicity; disrupting the helix at the base of the linker would result in the transition to more of a random coil. This notion is further supported by the data showing that interrupting the helix by introducing a separating proline disrupts the propagation of the helix into the linker, resulting in an unstructured linker.

While this chapter has primarily focused on the first 6 residues of the linker, it is notable that the linker ends with two proline residues. Although prolines are known to stably cap helices from N-terminal positions(Presta, Rose 1988), they have the opposite effect when located at the C-terminus and generally act to disrupt upstream helices(Kim, Kang 1999), and two consecutive proline residues are likely to even more strongly oppose helix formation. Overall, it seems possible then, that one or two of the linker residues preceding the two proline residues at the end of the AraC linker are not truly helical. Such a structure, appears to be compatible with all of the experiments reported here.

How could the structural state of the linker be controlled? As the linker is not an essential part of either domain and likely extends in space away from both of them, the only source of a signal to the linker is either from the N-terminal arm of AraC or directly through the residues of the dimerization domain immediately preceding the linker. Because the linker functions properly with a wide variety of sequences(Seedorff, Schleif 2011), a direct linker-arm interaction seems unlikely. The dimerizing helix immediately precedes the linker, and thus can nucleate helix formation of the linker. If arabinose controlled the helical stability or some other structural property of several of the residues of the dimerizing helix, the helical state of most of the linker could be controlled. This is an appealing mechanism because not only do

repression negative mutations lie in the N-terminal arm, they also lie in the portion of the protein between the apo position of the arm and the base of the linker (Figure 5). Additionally, this is consistent with previous experiments on the anisotropy of the dimerization domain(Malaga, Mayberry et al. 2016). Thus, a potential mechanism for communication of the presence of arabinose to the linkers could be through structural changes induced by the relocation of the N-terminal arm of AraC over arabinose that then allow the residues lying between the arm and the interdomain linker to alter their conformation, and thereby weaken the helix nucleating capability of the dimerizing helix.

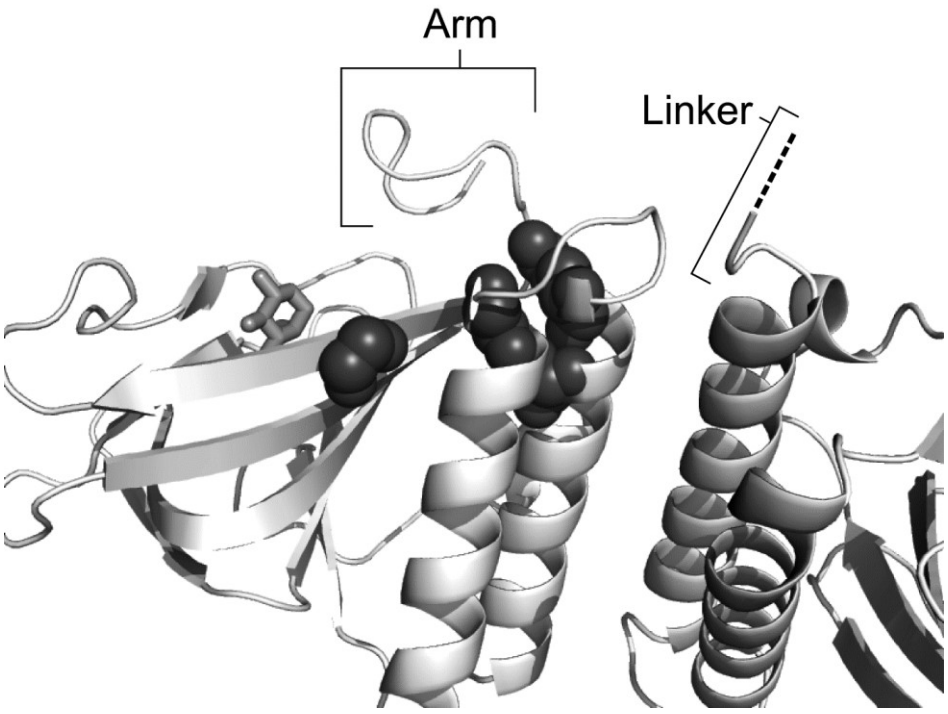

Figure 5. The positions of repression-disrupting mutations, G22, G141, A152, and E149, which are represented in space filling mode, in the region between arm and linker. Apo form protein is presented with arabinose (in stick form) superimposed to illustrate its binding location.

Materials and Methods

Plasmids and strains

Except where noted, mutants of AraC were generated using Stratagene QuikChange™ protocol for site-directed mutagenesis (Stratagene, La Jolla, CA) from oligonucleotide primers and verified by sequencing the products. The AraC variants used here additionally contained the Y31V mutation(Weldon, Rodgers et al. 2007), which reduces aggregation and facilitates *in vitro* experiments.

For the *in vivo* activity measurements of the 6A, GS linker variants, and the arm deletion, AraC the protein was expressed from the pWR03 AraC expression vector based on pSE380(Reed, Schleif 1999). For measurement of activity from AraC variants carried on this plasmid, it was transformed into a strain deleted of AraC, but containing *ara* p_{BAD} and the *araBAD* genes, thereby allowing measurement of arabinose isomerase for quantitation of the p_{BAD} basal levels. This strain, RS321, is derived from SH321(Hahn, Dunn et al. 1984) by P1 transduction transferring Tn10(laci^{Q} Cm^{r} at the *lac* locus) from Gottesman strain ASP7020 (NM514)(Parker, Gottesman 2016).

DNA used in binding measurements

The I_1-I_1 oligomers used in the DNA binding assays were 5'-/Cy5/gccaTAGCATTTTTATCCATAagatTAGCATTTTTATCCATAcctc-3' as well as the unlabeled reverse compliment. Underlined regions indicate I_1 binding sites. Specific competitor was the same sequence, but lacked the Cy5 label. Labeled DNA was annealed in 25 mM Tris pH 8, 1 mM EDTA, 2.5 mM MgCl, and 100 mM KCl at a 1:1.1 molar ratio of labeled:unlabeled strands in order to ensure that no fluorescent label remained single-stranded. Unlabeled competitor was annealed at an equimolar ratio of both strands. Annealing was performed by incubating at 95°C for 5 minutes before gradually reducing the temperature to 20°C over a period of 30 minutes.

Protein used in DNA binding measurements

For the *in vitro* DNA binding measurements, AraC was overproduced using the pET-24 expression vectors(Novagen) with AraC cloned between the *Nco*I and *Sac*I restriction sites, and isolated by ion exchange chromatography sequentially on heparin, HiTrap-Q, and again heparin columns as described by Rodgers and Schleif(Rodgers, Schleif 2012).

Arabinose isomerase assay

Arabinose isomerase assays were performed as described by Schleif and Wensink(Schleif, Wensink 1981). Cells, containing the plasmid, SH321 (*F⁻ ΔaraC-leu1022 Δlac74 galK⁻ Str^r thi⁻1⁴*) or SH322 (*F⁻ Δlac74 galK⁻ Str^r thi⁻1⁴*)(Hahn, Dunn et al. 1984) were grown in M10 minimal salts medium(Schleif, Wensink 1981) plus 0.4% glycerol or 0.4% arabinose, 10 µg/ml thiamine, 0.2% casamino acids, and 100 µg/mL ampicillin, to an apparent OD_{650} of 0.6 to 0.8 (as measured by a Perkin/Elmer Lambda 25 UV/Vis spectrophotometer) before harvesting and concentration for assay of isomerase. Cells containing pCCNC were grown to stationary phase in YT medium containing 40 µg/ml kanamycin for measurement of uninduced levels of arabinose isomerase.

DNA dissociation and electrophoretic mobility shift assay

The binding-dissociation experiments were done at 37° C in 10 mM Tris pH 7.4, 1 mM K-EDTA, 1 mM DTT, 5% glycerol, and 0.1 mg/mL BSA containing either 100 mM KCl for the minus-arabinose measurements or 300 mM KCl and 5 mM arabinose for the plus arabinose experiments. These concentrations were chosen to provide an optimum time scale for experimentation, allowing for sufficient resolution while removing the complications introduced by longer time courses. DNA and protein were equilibrated together using 50 nM labeled I_1-I_1 DNA, and 150 nM purified AraC protein in 10 µL volumes for 5 minutes prior to the addition of a 30 times excess of unlabeled I_1-I_1 DNA and 0.5 µg sheared calf thymus DNA in 10 µL. Samples were then loaded and immediately run on 6% acrylamide, 0.1% methylene bis-acrylamide gels containing 10 mM Tris-Acetate pH 7.4, 1 mM K-EDTA at 8 V/cm for 30 minutes.

Control experiments showed that DNA entered gels within 30 seconds of loading. Gels were imaged on a GE Typhoon 9410 fluorescence scanner using a 633 nm laser for excitation and 670 nm emission filter. The resultant images were analyzed using GE ImageQuant™ TL 7.0 software.

Circular dichroism (CD) measurements and sample preparation

N-terminal acetylated peptides for CD measurements were obtained from GenScript USA Inc. as a crude preparation, and further purified by reverse phase liquid chromatography using a PepRPC™ 15 μm column, eluting with acetonitrile, then dried under vacuum by centrifugal evaporation. CD spectra were taken on an Aviv 2.86 spectropolarimeter. Samples were prepared in 10 mM phosphate buffer pH 7, and 100 mM NaCl. Spectra were measured at 2.0° C at 1 nm intervals from 250 nm to 185 nm wavelengths, averaging over 5 seconds. Dynode exceeded 350 V at wavelengths below 190 nm, causing considerable uncertainty in this region. As such, these measurements were discounted during analysis. To obtain the reported values, CD signal from buffer alone was subtracted from all spectra, after which values were converted into mean residue ellipticity units.

References

BERG, O. and VON HIPPEL, P.H., 1985. Diffusion-controlled macromolecular interactions. *Annu Rev Biophys Biophys Chem,* **14**, pp. 131-160.

BRAHMS, S. and BRAHMS, J., 1980. Determination of protein secondary structure in solution by vacuum ultraviolet circular dichroism. *Journal of Molecular Biology,* **138**(2), pp. 149-178.

BUSTOS, S.A. and SCHLEIF, R.F., 1993. Functional domains of the AraC protein. *Proceedings of the National Academy of Sciences of the United States of America,* **90**(12), pp. 5638-5642.

CARRA, J.H. and SCHLEIF, R.F., 1993. Variation of half-site organization and DNA looping by AraC protein. *The EMBO journal,* **12**(1), pp. 35-44.

DUNN, T.M., HAHN, S., OGDEN, S. and SCHLEIF, R.F., 1984. An operator at -280 base pairs that is required for repression of *araBAD* operon promoter: addition of DNA helical turns between the operator and promoter cyclically hinders repression. *Proceedings of the National Academy of Sciences of the United States of America,* **81**(16), pp. 5017-5020.

GUTFREUND, H., 1995. *Kinetics for the Life Sciences.* Cambridge, UK: Cambridge University Press.

HAHN, S., DUNN, T. and SCHLEIF, R., 1984. Upstream repression and CRP stimulation of the *Escherichia coli* L-arabinose operon. *Journal of Molecular Biology,* **180**(1), pp. 61-72.

HAHN, S. and SCHLEIF, R., 1983. *In vivo* regulation of the *Escherichia coli araC* promoter. *Journal of Bacteriology,* **155**(2), pp. 593-600.

HENDRICKSON, W. and SCHLEIF, R., 1985. A dimer of AraC protein contacts three adjacent major groove regions of the *araI* DNA site. *Proceedings of the National Academy of Sciences of the United States of America,* **82**(10), pp. 3129-3133.

HENDRICKSON, W. and SCHLEIF, R.F., 1984. Regulation of the *Escherichia coli* L-arabinose operon studied by gel electrophoresis DNA binding assay. *Journal of Molecular Biology,* **178**(3), pp. 611-628.

KIM, M.K. and KANG, Y.K., 1999. Positional preference of proline in α-helices. *Protein Science,* **8**(7), pp. 1492-1499.

KUKIC, P., CAMILLONI, C., CAVALLI, A. and VENDRUSCOLO, M., 2014. Determination of the Individual Roles of the Linker Residues in the Interdomain Motions of Calmodulin Using NMR Chemical Shifts. *Journal of Molecular Biology,* **426**(8), pp. 1826-1838.

LOBELL, R.B. and SCHLEIF, R.F., 1990. DNA looping and unlooping by AraC protein. *Science (New York, N.Y.),* **250**(4980), pp. 528-532.

MA, B., TSAI, C., HALILOĞLU, T. and NUSSINOV, R., 2011. Dynamic Allostery: Linkers Are Not Merely Flexible. *Structure,* **19**(7), pp. 907-917.

MALAGA, F., MAYBERRY, O., PARK, D., RODGERS, M.E., TOPTYGIN, D. and SCHLEIF, R.F., 2016. A genetic and physical study of the interdomain linker of E. Coli AraC protein-a trans-subunit communication pathway. *Proteins: Structure, Function, and Bioinformatics,* **84**, pp. 448--460.

MARQUSEE, S., ROBBINS, V.H. and BALDWIN, R.L., 1989. Unusually stable helix formation in short alanine-based peptides. *Proceedings of the National Academy of Sciences of the United States of America,* **86**(14), pp. 5286-5290.

MARTIN, K., HUO, L. and SCHLEIF, R.F., 1986. The DNA loop model for *ara* repression: AraC protein occupies the proposed loop sites *in vivo* and repression-negative mutations lie in these same sites. *Proceedings of the National Academy of Sciences of the United States of America,* **83**(11), pp. 3654-3658.

PACE, C.N. and SCHOLTZ, J.M., 1998. A Helix Propensity Scale Based on Experimental Studies of Peptides and Proteins. *Biophysical journal,* **75**(1), pp. 422-427.

PARKER, A. and GOTTESMAN, S., 2016. Small RNA Regulation of TolC, the Outer Membrane Component of Bacterial Multidrug Transporters. *J. Bacteriol.,* **198**(7), pp. 1101-1113.

PRESTA, L.G. and ROSE, G.D., 1988. Helix signals in proteins. *Science,* **240**(4859), pp. 1632.

REED, W.L. and SCHLEIF, R.F., 1999. Hemiplegic mutations in AraC protein. *Journal of Molecular Biology,* **294**(2), pp. 417-425.

RODGERS, M.E. and SCHLEIF, R., 2009. Solution structure of the DNA binding domain of AraC protein. *Proteins,* **77**(1), pp. 202-208.

RODGERS, M.E. and SCHLEIF, R., 2012. Heterodimers reveal that two arabinose molecules are required for the normal arabinose response of AraC. *Biochemistry,* **51**(41), pp. 8085-8091.

SCHLEIF, R.F. and WENSINK, P.C., 1981. *Practical methods in molecular biology.* New York: Springer-Verlag.

SEABOLD, R.R. and SCHLEIF, R.F., 1998. Apo-AraC actively seeks to loop. *Journal of Molecular Biology,* **278**(3), pp. 529-538.

SEEDORFF, J. and SCHLEIF, R., 2011. Active role of the interdomain linker of AraC. *Journal of Bacteriology,* **193**(20), pp. 5737-5746.

SHEPPARD, D.E. and ENGLESBERG, E., 1967. Further evidence for positive control of the L-arabinose system by gene *araC*. *Journal of Molecular Biology,* **25**, pp. 443-454.

SOISSON, S.M., MACDOUGALL-SHACKLETON, B., SCHLEIF, R. and WOLBERGER, C., 1997. The 1.6 A crystal structure of the AraC sugar-binding and dimerization domain complexed with D-fucose. *Journal of Molecular Biology,* **273**(1), pp. 226-237.

WANG, X., VALLURUPALLI, P., VU, A., LEE, K., SUN, S., BAI, W., WU, C., ZHOU, H., SHEA, J., KAY, L.E. and DAHLQUIST, F.W., 2014. The Linker between the Dimerization and Catalytic Domains of the CheA Histidine Kinase Propagates Changes in Structure and Dynamics That Are Important for Enzymatic Activity. *Biochemistry,* **53**(5), pp. 855-861.

WELDON, J.E., RODGERS, M.E., LARKIN, C. and SCHLEIF, R.F., 2007. Structure and properties of a truely apo form of AraC dimerization domain. *Proteins,* **66**(3), pp. 646-654.

WRIGGERS, W., CHAKRAVARTY, S. and JENNINGS, P.A., 2005. Control of protein functional dynamics by peptide linkers. *Biopolymers,* **80**(6), pp. 736-746.

ZHANG, X., REEDER, T. and SCHLEIF, R., 1996. Transcription activation parameters at *ara* p_{BAD}. *Journal of Molecular Biology,* **258**(1), pp. 14-24.

Chapter 4: Residue covariance identifies interdomain contacts

Abstract

E. coli AraC protein regulates expression of the L-arabinose operon by binding DNA upstream of the p_{BAD} promoter. The apo form of AraC orients its DNA-binding domains (DBDs) so as to bind to DNA half-sites 210 base-pairs apart, looping DNA and repressing expression of the operon. The domains must reorient when bound to arabinose, contacting adjacent half-sites instead allowing for induction. Past experiments have established that the separate DNA-binding- and dimerization domains (DDs) must interact, likely helping to stabilize the inducing and/or repressing states, though the specific locations of these interactions have not been identified. Here I employ a bioinformatic approach, identifying covarying residues within a large collection of homologs, with the underlying assumption that strongly covarying residues exist in close proximity. I combine this with classical screening methods to locate residue pairs involved in these interdomain contacts. These findings suggest a structural model, consistent with previous findings, of a DD-DBD interface during induction, stabilizing the holo form of the protein allowing for proper induction.

Introduction

AraC must undergo a structural change in order to bind alternately to two different sets of DNA binding sites (distant sites in the absence of arabinose(Dunn, Hahn et al. 1984) and directly repeating sites in its presence(Carra, Schleif 1993, Martin, Huo et al. 1986)). In previous chapters I have shown that this change is likely propagated from a structural change in the N-terminal arm through nearby residues (Chapter 2), leading to a structural change of the interdomain linker (Chapter 3). The logical next question is, how does this linker change result in an altered affinity for DNA biding sites. To phrase it more directly, "what happens next?" The answer to this question likely falls into one of two different possibilities: either the structural signal from the linker continues to propagate through the DBD causing an altered affinity for specific sequences, or the orientations of the DBDs are changed in such a manner as to promote binding to

DNA sites in particular orientations. The former possibility – that of altered sequence affinity – can promptly be ruled out however, as the protein has shown consistent DNA sequence affinity in both the presence and absence of arabinose(Hendrickson, Schleif 1984, Carra, Schleif 1993). Thus, the structural change must involve an altered preference of the protein for adjacent sites, and the aforementioned change in the linker likely results a larger-scale rearrangement of the domains.

Is the linker the whole story, or are further interactions stabilizing the inducing form of the protein? Previous experiments have proven that interactions between the dimerization domains and the DBDs do, in fact, exist. Free DBD displays different binding affinity for its DNA sites compared to full-length AraC(Cole, Schleif 2012), indicating that such interactions must exist, and a DNA-assisted binding assay identified the existence of such an interaction between domains(Frato, Schleif 2009), though the exact nature has remained mysterious.

Unfortunately, putative inter-domain contacts in AraC cannot be identified by direct structural examination. The crystal structures of both the DBD(Rodgers, Schleif 2009) and the dimerization domain (in both the presence(Soisson, MacDougall-Shackleton et al. 1997) and absence(Weldon, Rodgers et al. 2007) of arabinose) have been solved, but despite many attempts, no full-length structure AraC has been directly ascertained. Decades of efforts have failed to crystalize the full-length protein in either the presence or absence of ligand, and more recent attempts at structural resolution via cryo-electron microscopy have proven unfeasible. Thus, less direct methods are currently being employed to determine the nature of the alternate structures.

One of the most typical classical method for identifying functional residue-residue contacts is a simple second-site suppressor screen, whereby a phenotype caused by a mutation in one residue is rescued by a corresponding mutation in a different one. This operates under the assumption that contacts altered by the change in one residue can be recovered by corresponding changes in the contacted one. While past experiments have identified a small number of possible candidates(Ross, Gryczynski et al. 2003, Dirla, Chien et al. 2009), no plausible functional relationships have been established. The combinatorics of

screening every potential residue are staggering, and truly beyond the scope of a reasonable experiment of this nature. However, recent advances have made it possible to co-opt nature's own second-site suppressor screens by examining evolutionary amino acid covariance.

Nucleic acid covariance has long been utilized for identifying regions of interaction, particularly in the RNA community(Eddy, Durbin 1994). Nucleotides involved in Watson-Crick base-pairing must compensate for a single nucleotide polymorphism with a corresponding one in order to maintain structure and restore function. Thus, functional nucleotide interactions can be identified by discovering nucleotide pairs that recover in this predictable manner (Figure 1). Amino acid residue interactions, however are far more complicated and less predictable; unfortunately, there are no equivalent pairwise "rules" for effective residue-residue contacts as there are for nucleotides. However, the same fundamental concept is conserved: a change in one residue that alters the shape of the interface can be rescued by a complementary change in a contacting residue, so residue pairs that change in concert with one another, throughout a list of homologs, are likely to interact.

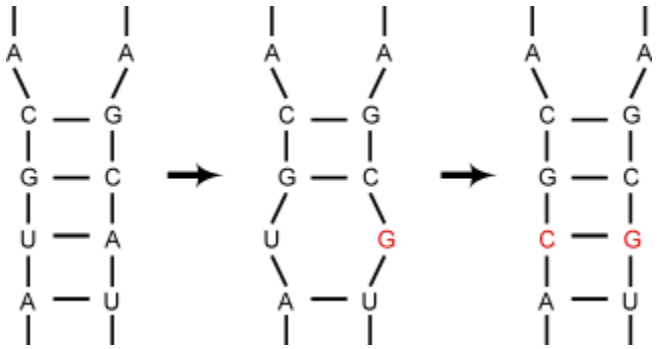

Figure 1. Example of identification of nucleotide interaction by covariance.

Using bioinformatic tools, it is possible to analyze a large set of AraC homologs for residue covariances. The GREMLIN software, developed by the Baker lab at the University of Washington, allows for the comparison of thousands of homologous sequences and identification of covarying residue pairs(Kamisetty, Ovchinnikov et al. 2013, Ovchinnikov, Kamisetty et al. 2014). Using this software, I was able to identify several covarying residue pairs linked across the different domains of the protein. I was then

able to test these suggested residue pairs directly via site-directed mutagenesis, identifying pairs of mutations that individually abolish inducibility, but rescue activity when present together.

Results

Generation of input dataset

To generate a list of functional AraC homologs, I generated an alignment beginning with sequences obtained from a large (10,000-sequence) alignment of AraC homologs produced by a BLAST search of the canonical *E. coli* AraC sequence. These 10,000 initial sequences were pared down to 987 non-redundant sequences. The AraC family of proteins comprises a wide range of homologous molecules, many of which possess functions distinct from AraC's specific role of *ara* operon-regulation(Gallegos, Schleif et al. 1997); such functionally-divergent homologs almost certainly evolved under different functional and structural constraints, making their inclusion potentially confounding to my effort of identifying covariance meaningful for AraC function. Thus, I subsequently manually curated this initial list, ensuring that all representative sequences of an arbitrary sample of approximately 50 sequences – each from a different bacterial species – localized to the L-arabinose operon of the host bacteria.

The GREMLIN software automatically and unavoidably discards sequence regions below a certain threshold of coverage/gaps. Unfortunately, the functionally crucial N-terminal arm is such a gap-filled region, meaning that initial runs of GREMLIN discarded the arm entirely from consideration. Since (as explored in Chapter 2) the arm is a plausible region of interdomain contact, I further edited the alignment to force these early residues into the data. In order to accomplish this without sacrificing the validity of the data, I designed a script so replace any gaps in the multiple alignment text file with a residue chosen at random. Additionally, because the first step of the GREMLIN computation is to enrich the alignment with additional homologous sequences found by GREMLIN, my gap-filled alignment solved the additional problem of the software potentially inserting any of the above-mentioned functionally-divergent AraC homologs, as the overall sequences were so significantly altered. I generated three separate randomly-gap-

filled alignment files to eliminate the risk of false covariances being inadvertently generated by the randomly generated residues

Prediction of functional contacts

To generate a list of putatively interacting, covarying residues, I input each of the three gap-filled alignments, separately, into the GREMLIN software, generating three separate lists of predictions. Any pairs not present in all three datasets were discarded, and the rest condensed into a single list. The GREMLIN software has previously been verified as enriching closely-associating residues in other proteins(Kamisetty, Ovchinnikov et al. 2013, Ovchinnikov, Kamisetty et al. 2014), but, as a proof-of-concept to ensure that the system accurately functioned for AraC, I verified that the intra-domain contacts predicted agreed with the known crystal structures of the protein domains. As shown in Figure 2, the GREMLIN-generated intra-domain pairs are greatly enriched in proximate residues compared to a data set of all inter-residue distances in the structure, indicating that the system indeed generates plausible contact predictions. Additionally, GREMLIN's automatic analysis identified the PDB crystal structure of *E. coli* AraC dimerization domain as the most closely matching structure to that derived from the GREMLIN prediction.

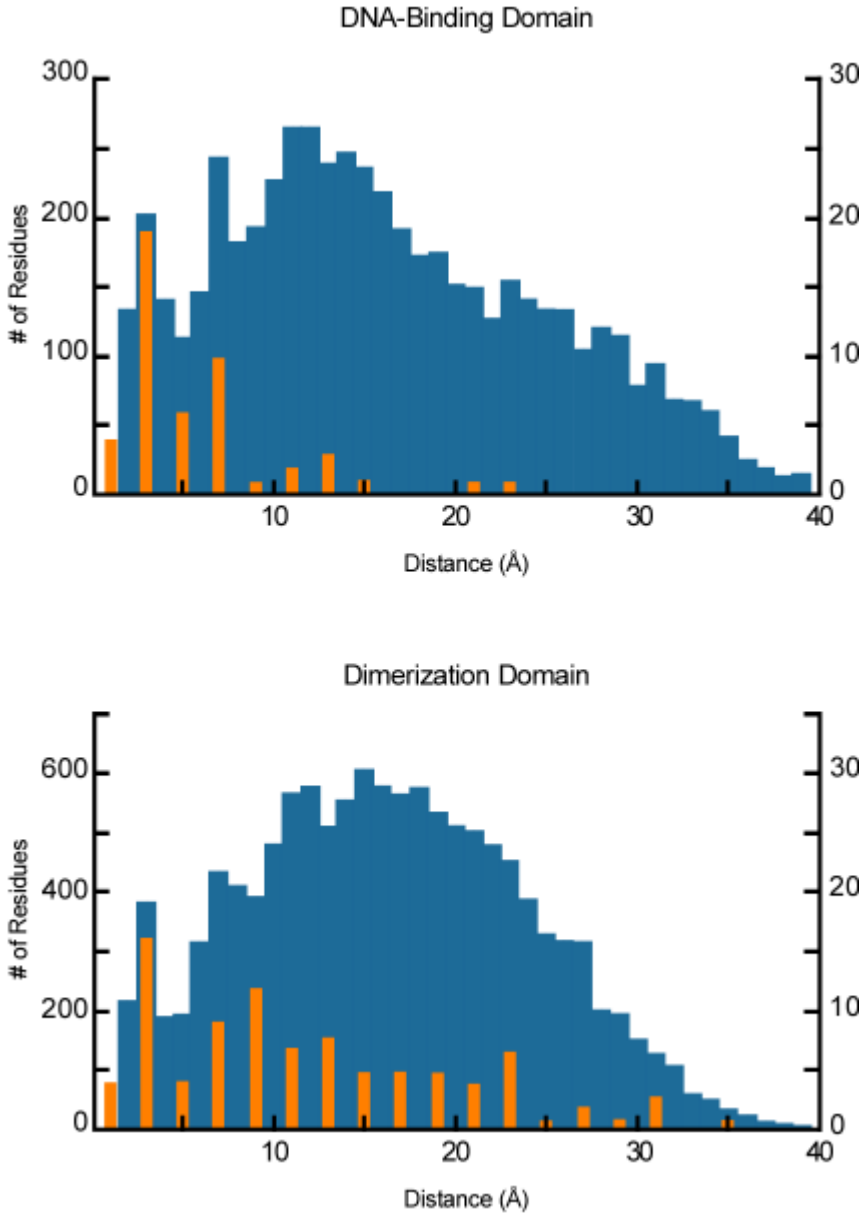

Figure 2. Enrichment of nearby residues in GREMLIN-predicted pairs. Histograms of distances of all GREMLIN predicted intra-domain pairs (orange, right axis), compared to all inter-residue distances in the crystal structures (blue, left axis). Distances were measured from the center-of-mass of each residue. DNA-binding domain structure (top): RCSB PDB file 2k9s. Dimerization domain structure (bottom): RCSB PDB file 2arc

In addition to echoing proximity data from the published crystal structures(Soisson, MacDougall-Shackleton et al. 1997, Weldon, Rodgers et al. 2007, Rodgers, Schleif 2009), the GREMLIN output notably predicted inter-domain contacts between the surface of the DBD and the central region of the dimerization domains; i.e. near the dimerization interface, as shown in Figure 3. These locations are consistent with previous models that were based primarily on the length of the interdomain linkers and the spacing of the DNA binding sites, allowing one DBD to lie over this central region of the protein during induction. Since the DBD is approximately the same length as the dimerization domain (along the axis of its dimerization interface), this model allows one DBD to "double back" across the protein, while the other bends to contact DNA in its proximity. Virtually any other structure would require moderate to significant unfolding of the protein. It is also worth noting that GREMLIN predicted contacts between pairs of residues in the N-terminal arm, as well as between the arm and other residues in the dimerization domain, but it identified no covariation between any arm residues and any DBD residues. This further reinforces findings previously discussed in Chapter 2 regarding the lack of direct interaction between the arm and DBD.

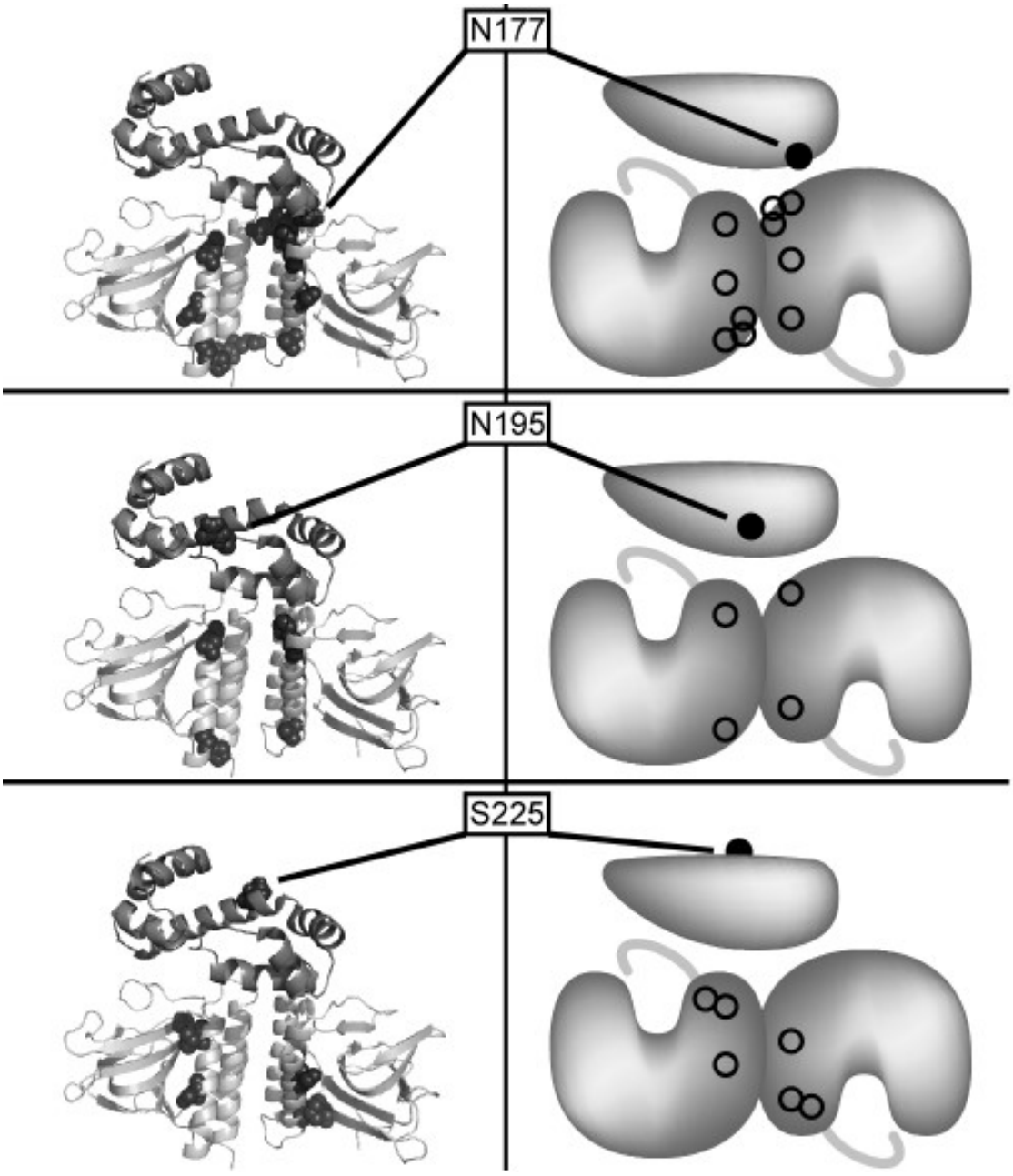

Figure 3. GREMLIN-predicted locations of possible inter-domain contacts in AraC. Closed circles represent DBD residues identified as covarying with each of the corresponding open circles on dimerization domain. One DBD is shown; the top surface represents the DNA-binding face

Effect of variation between predicted residue pairs

To test the validity of the GREMLIN-predicted inter-domain contacts, I screened for rescues among the predicted pairs. Figure 3 and Table 1 show the locations of the predicted pairs. In each case, the relevant

DBD residue first was selectively mutagenized, randomizing the codon to produce a full array of point mutants. This random pool was then screened by examining fluorescence of GFP under the control of the p_{BAD} promoter and full upstream regulatory region. Colonies were selected for lack of inducibility – that is, non-fluorescence on arabinose-containing plates, their plasmids isolated and pooled. These non-inducing mutants were then mutagenized again, altering one of the corresponding predicted dimerization domain residues, again at random. These double-mutants were again plated, and screened for fluorescent (rescued) colonies. In all cases, all colonies displayed either very strong fluorescence or complete lack of fluorescence, and no intermediate phenotypes were observed. Given this binary output, fluorescence intensity was not quantified.

Table 1 GREMLIN-predicted DD-DBD contact locations

DBD residue[a]	DD residue[a]	Predicted probability[b]
N177	A166	0.95
	N139	0.83
	H126	0.82
	E165	0.81
	S131	0.73
N194	N139	0.90
	H126	0.76
S225	I138	0.90
	A21	0.76
	S131	0.67

[a]residue designations reported as *E. coli* identities
[b]Probability values calculated by GREMLIN software for given pairwise contact

GREMLIN predicted three DBD residues, above a probability threshold of 0.7, involved in interdomain contacts with the DD. All three DBD residues listed among the predictions indeed produced mutants incapable of inducing on arabinose-rich media. However, of the DD residues tested, none of the predicted contacts with S225 produced mutants capable of rescuing inducibility. However, some N194 and N177 non-inducing mutants were capable of being rescued by mutations in residues N139 and H126 (Table 2). However, after sequencing, I discovered that the rescued N177 mutants all contained insertions at that site, which is especially noteworthy because N177 is located within an alpha helix according to the crystal structure(Soisson, MacDougall-Shackleton et al. 1997, Weldon, Rodgers et al. 2007), suggesting that an insertion at this location could alter the position of the downstream residues, including N194. Thus, the data suggest that N194 on the DBD contacts N139 and H126 on the dimerization domain during induction. As can be seen in Figure 3, these two residues are appropriately close to each other on opposite subunits (12.3Å) for a contacting residue to fit between.

Table 2 Successful DD-DBD contact rescues

DBD residue N194-	DD residue Rescue[a]
G	N139A
	H126P
I	N139L

[a]DD residues listed were shown to rescue associated N194 mutant

Discussion

The work described here identified previously unknown residue change pairs that restored normal AraC activity. The pairs involved are therefore assumed to be in necessary contact for the inducing form of

AraC. This reinforces previous evidence of interaction between the dimerization and DNA-binding domains(Cole, Schleif 2012, Frato, Schleif 2009). Using the GREMLIN software, I was able to use the natural evolution of AraC throughout the prokaryotic domain to identify covarying residues within the protein, streamlining the process of suppressor screening, allowing for an alternate manner of protein structure characterization where full-length structural data was unavailable. This covariation analysis, refined through direct mutagenesis of the putative reported contacts, suggests a particular structure of AraC in the presence of arabinose.

The GREMLIN software reported three main sites of the DBD possibly involved in interdomain contact. Ultimately, I believe that the most relevant of these was N194, which mutagenesis experiments showed to interact with H126 and N139 on the dimerization domain. These two dimerization domain residues are both located on the first helix of the antiparallel coiled-coil of the dimerization interface, and though they are located toward opposite ends of this helix, they align directly on opposite subunits (Figure 3), creating a highly plausible site for interaction. As shown in Figure 4, the sidechains if H126 and N139 in fact are oriented towards one another, and N194 on the DBD can insert directly between these two. Additionally, this arrangement of domains leaves the DNA-binding surface of the DBD oriented outwards and free to bind to its target, as would be a requirement for any model. Furthermore, were the protein to adopt this conformation, with this DBD lying across the dimerization domains and bound to I_1, the second DBD would become positioned directly above the I_2 site, favoring easy binding to it.

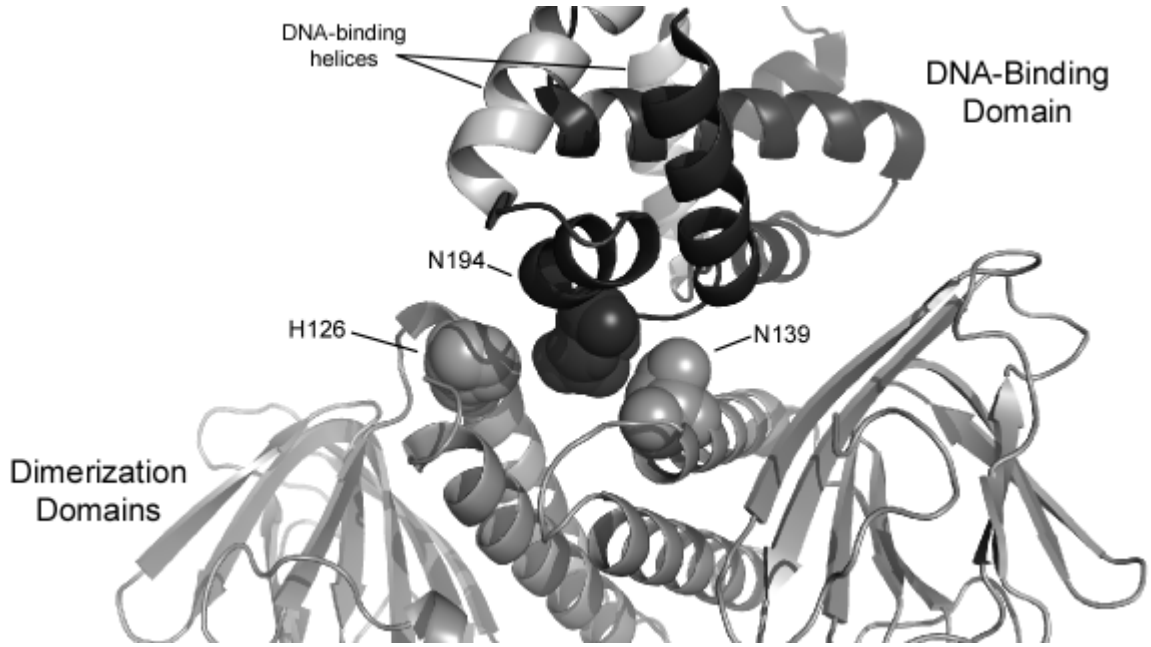

Figure 4. Potential model of DBD-DD interaction

Of the three DBD residues identified by GREMLIN, one residue, S225, is located directly on the DNA-binding face of the DBD, which cast significant immediate doubt that it could be involved in direct contact with the dimerization domain. S225 falls between the two DNA-binding helices, leaving no plausible way for this residue to contact other regions while the protein is bound to DNA. As expected, mutagenesis experiments identified no second-site suppressors for S225 mutants among the GREMLIN predicted pairs. Curiously, one particular dimerization domain residue, I138, seems to covary quite strongly with S225. This could suggest an evolutionarily divergent function among a subset of AraC homologs, or a possible allosteric pathway involving the two. However, random chance cannot be ruled out as the force behind this particular covariance.

Although both H126 and N139 showed both covariance and rescue of N177 in addition to N194, I do not believe this represents direct interaction. The N177 mutations rescued by these dimerization domain residues were all insertions. That is, all single substitutions of N177 on the DBD were un-rescuable by the dimerization domain mutations tested, and rescue only occurred when the DBD mutation resulted in an

additional residue. N177 is found towards the beginning of the entry helix of the DBD, so it is plausible that an insertion would result in either an elongation or disruption of this helix. This would cause the end of the helix to be positioned differently, affecting the orientations of downstream residues. In fact, N194 itself is found just beyond the end of this helix, so a change in this helix would affect its position, still necessitating an alteration in any residues it interacts with. Thus, these findings are all consistent with a mechanism involving N194 contacting the dimerization domain.

Materials and Methods

Identification of AraC homologs

AraC homologs were obtained using NCBI protein BLAST software aligning non-*Escherichia* sequences to the canonical *E. coli* AraC protein sequence, requesting 10,000 homologous sequences. An arbitrary set of 50 putative orthologs were selected and examined using the NCBI gene viewer to ensure localization within the L-arabinose operon regulatory region of the genome.

Analysis of covariance using GREMLIN software

To ensure analysis across the full length of the aligned sequences, a Python script was designed to replace each gap character with a randomly-generated amino acid character. Final alignment of 987(x3) sequences was submitted to GREMLIN software (http://gremlin.bakerlab.org/index.php), and the resultant lists were aggregated, condensed, and sorted using Microsoft Excel™. A cutoff S score of 1.174, equating to a reported probability of 0.7, was used to select residue pairs to be tested.

Generation of AraC mutants

AraC mutants were generated using Stratagene QuikChange™ protocol for site-directed mutagenesis (Stratagene, La Jolla, CA) from oligonucleotide primers with the target codon altered to random trinucleotide sequences. Non-inducing DBD mutant colonies were collected and combined before plasmid extraction, and all mutants of a single DBD residue were secondarily mutagenized in a single reaction. Final rescue double-mutants were sequenced.

p_{BAD}-GFP fluorescence assay

AraC protein was expressed from the variable copy number plasmid pCCN AraC expression vector based on the p_{BAD}-GFP plasmid. This variable copy number plasmid contains two origins of replication: f1 which will maintain a single copy per cell and oriS which allows for high copy-number but is blocked by LacI if present at elevated levels. it was transformed into a strain deleted of AraC, but containing *ara* p_{BAD} and the *araBAD* genes, thereby allowing measurement of GFP expression without interference of genomic AraC. This strain, RS321, is derived from SH321 (F^- *ΔaraC-leu1022 Δlac74 galK*$^-$ *Str*^{r} *thi*$^-$*1*^{4})(Hahn, Dunn et al. 1984) by P1 transduction transferring Tn10(laci^Q Cm^r at the *lac* locus) from Gottesman strain ASP7020 (NM514)(Parker, Gottesman 2016).

Cells were plated on YT plates containing 0.4% arabinose and visually examined under UV light. All nonfluorescent colonies were picked and restreaked on new plates to remove possible contamination from nearby colonies. Still nonfluorescent colonies were pooled and mutagenized as described above before transforming and plating. All resultant double-mutagenized colonies were spotted onto new plates for fluorescence analysis. As no intermediate phenotypes were observed, fluorescence intensity was not quantified, and fluorescent/nonfluorescent phenotypes were evaluated qualitatively by eye.

References

CARRA, J.H. and SCHLEIF, R.F., 1993. Variation of half-site organization and DNA looping by AraC protein. *The EMBO journal,* **12**(1), pp. 35-44.

COLE, S.D. and SCHLEIF, R., 2012. A new and unexpected domain-domain interaction in the AraC protein. *Proteins: Structure, Function, and Bioinformatics,* **80**(5), pp. 1465-1475.

DIRLA, S., CHIEN, J.Y. and SCHLEIF, R., 2009. Constitutive mutations in the *Escherichia coli* AraC protein. *Journal of Bacteriology,* **191**(8), pp. 2668-2674.

DUNN, T.M., HAHN, S., OGDEN, S. and SCHLEIF, R.F., 1984. An operator at -280 base pairs that is required for repression of *araBAD* operon promoter: addition of DNA helical turns between the operator and promoter cyclically hinders repression. *Proceedings of the National Academy of Sciences of the United States of America,* **81**(16), pp. 5017-5020.

EDDY, S.R. and DURBIN, R., 1994. RNA sequence analysis using covariance models. *Nucleic acids research,* **22**(11), pp. 2079-2088.

FRATO, K.E. and SCHLEIF, R.F., 2009. A DNA-assisted binding assay for weak protein-protein interactions. *Journal of Molecular Biology,* **394**(5), pp. 805-814.

GALLEGOS, M.T., SCHLEIF, R., BAIROCH, A., HOFMANN, K. and RAMOS, J.L., 1997. Arac/XylS family of transcriptional regulators. *Microbiology and molecular biology reviews : MMBR,* **61**(4), pp. 393-410.

HAHN, S., DUNN, T. and SCHLEIF, R., 1984. Upstream repression and CRP stimulation of the *Escherichia coli* L-arabinose operon. *Journal of Molecular Biology,* **180**(1), pp. 61-72.

HENDRICKSON, W. and SCHLEIF, R.F., 1984. Regulation of the *Escherichia coli* L-arabinose operon studied by gel electrophoresis DNA binding assay. *Journal of Molecular Biology,* **178**(3), pp. 611-628.

KAMISETTY, H., OVCHINNIKOV, S. and BAKER, D., 2013. Assessing the utility of coevolution-based residueâ€“residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci USA,* **110**(39), pp. 15674.

MARTIN, K., HUO, L. and SCHLEIF, R.F., 1986. The DNA loop model for *ara* repression: AraC protein occupies the proposed loop sites *in vivo* and repression-negative mutations lie in these same sites. *Proceedings of the National Academy of Sciences of the United States of America,* **83**(11), pp. 3654-3658.

OVCHINNIKOV, S., KAMISETTY, H. and BAKER, D., 2014. Robust and accurate prediction of residueâ€“residue interactions across protein interfaces using evolutionary information. *eLife,* **3**, pp. e02030.

PARKER, A. and GOTTESMAN, S., 2016. Small RNA Regulation of TolC, the Outer Membrane Component of Bacterial Multidrug Transporters. *J. Bacteriol.,* **198**(7), pp. 1101-1113.

RODGERS, M.E. and SCHLEIF, R., 2009. Solution structure of the DNA binding domain of AraC protein. *Proteins,* **77**(1), pp. 202-208.

ROSS, J.J., GRYCZYNSKI, U. and SCHLEIF, R., 2003. Mutational analysis of residue roles in AraC function. *Journal of Molecular Biology,* **328**(1), pp. 85-93.

SOISSON, S.M., MACDOUGALL-SHACKLETON, B., SCHLEIF, R. and WOLBERGER, C., 1997. The 1.6 A crystal structure of the AraC sugar-binding and dimerization domain complexed with D-fucose. *Journal of Molecular Biology,* **273**(1), pp. 226-237.

WELDON, J.E., RODGERS, M.E., LARKIN, C. and SCHLEIF, R.F., 2007. Structure and properties of a truely apo form of AraC dimerization domain. *Proteins,* **66**(3), pp. 646-654.

Chapter 5: Conclusion/Perspectives – Where we have come to with AraC

Abstract

This thesis represents the culmination of decades-long investigations into the mechanisms of AraC. Previous work provided a general framework for the overall behavior of the protein and its behavior in regards to the L-arabinose operon in *E. coli*, while my work has revealed a probable pathway of intra-molecular signals constituting the overall mechanism for regulation of arabinose operon. This mechanism involves a change in the structure of the N-terminal arm causing subsequent repositioning of immediately adjacent regions of the protein propagating ultimately to the base of the interdomain linker. This can serve to break the helix directly preceding the linker, resulting in corresponding unfolding of the linker helix itself. This allows the newly flexiblized linker to reposition the DNA binding domains into an orientation stabilized by surface interactions between the DNA-binding and dimerization domains, allowing the protein to bind to adjacent DNA sites.

Discussion

I have presented here what is likely the final phase of the Schleif lab's work studying the L-arabinose operon, AraC protein, and the physical mechanism thereof. Previous work in the lab had determined the general behavior of the operon, and AraC's role in its regulation, though it is only now that an atomic-scale resolution of the mechanism has become clear. The work presented here, when taken as a whole, elucidates a credible allosteric signal pathway responding to arabinose binding. This signal begins with the N-terminal arm and proceeds through the dimerization domain core to alter the linker, which in-turn allows for new interdomain contacts that ultimately favor a protein structure capable of binding to adjacent DNA sites. In light of this new data, I will provide in this chapter a summary of the new proposed

mechanism for regulation of the L-arabinose operon by AraC. Considering the apo form of the protein as the initial reference state, the proposed allosteric mechanism unfolds in approximately 3 discrete steps (Figure 1).

Step one: Arm folds over arabinose

Chapter 2 of this dissertation presented evidence of functional intra-arm contacts governing repression. As noted, previous work had implicated the N-terminal arm as a critical component for repressibility, but failed to identify any residues participating in interdomain interactions (Ross, Gryczynski et al. 2003, Dirla, Chien et al. 2009). I have now shown evidence that the repressive interactions made by arm residues are in fact made with other arm residues. Thus, confirming previous speculations, repression is likely not achieved by the arm interacting directly with other regions, but rather it is the self-contained structure of the arm itself that causes AraC to favor its repressing state.

If we consider the apo form of AraC to be our reference state, the protein dimer begins with its N-terminal arms internally structured, and the protein's DNA-binding domains (DBDs) held away from the dimerization domain "body" of the protein by helical linkers. As previously shown, when arabinose enters the binding pocket of AraC, the nearby arm moves to bind over the sugar(Soisson, MacDougall-Shackleton et al. 1997, Weldon, Rodgers et al. 2007). This disrupts the previous structure of the arms. Given the close proximity of the arm and the sugar-binding pocket, no other signal should necessarily precede this arm response, making this the likely first step in the allosteric pathway.

As the arm shifts to bind over the ligand, the immediately following (in primary structure) residues must change position to accommodate this new location, as can be seen in an overlay of the apo(Weldon,

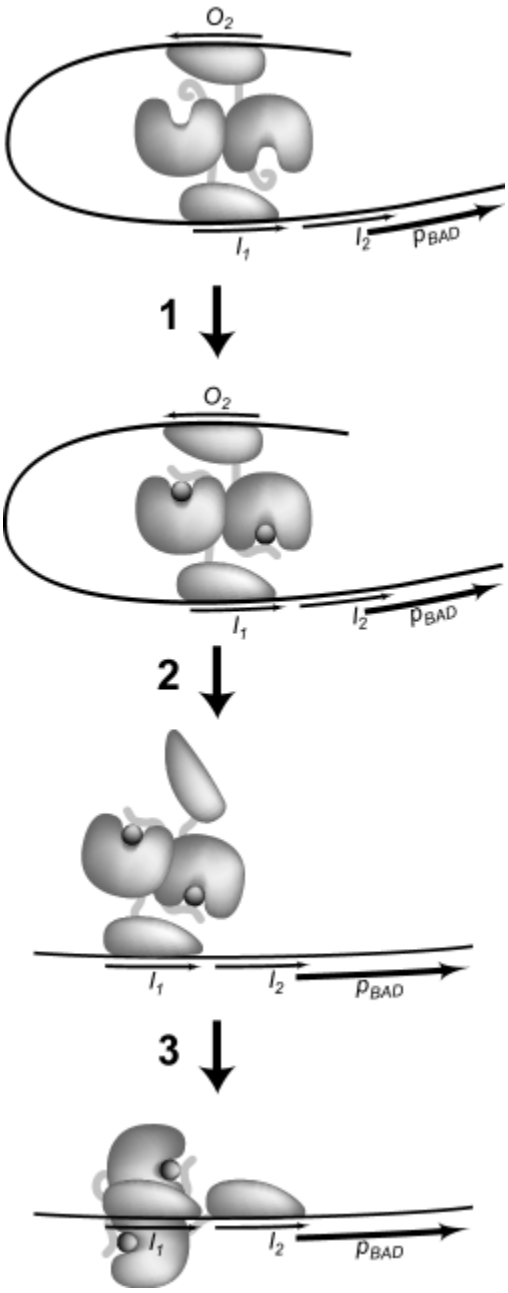

Figure 1. Mechanism of AraC. AraC transitions from repressing to inducing in 3 steps: (1) arm folds over arabinose. (2) linker transitions out of helical state. (3) interdomain contacts stabilize new conformation

Rodgers et al. 2007) and holo(Soisson, MacDougall-Shackleton et al. 1997) crystal structures of the dimerization domain (Figure 2). In both structures, residue A21 appears at the beginning of the first beta sheet of the arabinose-binding pocket, but the presence of arabinose causes this region to be pulled inward

toward the sugar and away from the nearby helices of the dimerization interface. Thus, in the holo form, this region can be thought of as existing in a tensed state, with the apo form representing the relaxed state. Though the sidechain of A21 is not shown in Figure 2, it is oriented toward and contacting the adjacent helix in the dimerization coiled-coil. Thus, this shift opens up a region in this area of the protein core, creating a resultant shift in the orientation of this helix, which in turn must propagate to the adjoining helix, and into the other subunit. This effect can be observed in Figure 2 by the progressive misalignment of successive helices, ultimately changing the environment at the end of the third helix (the C-terminal helix of the opposite subunit's dimerization domain), which also forms the base of that subunit's interdomain linker.

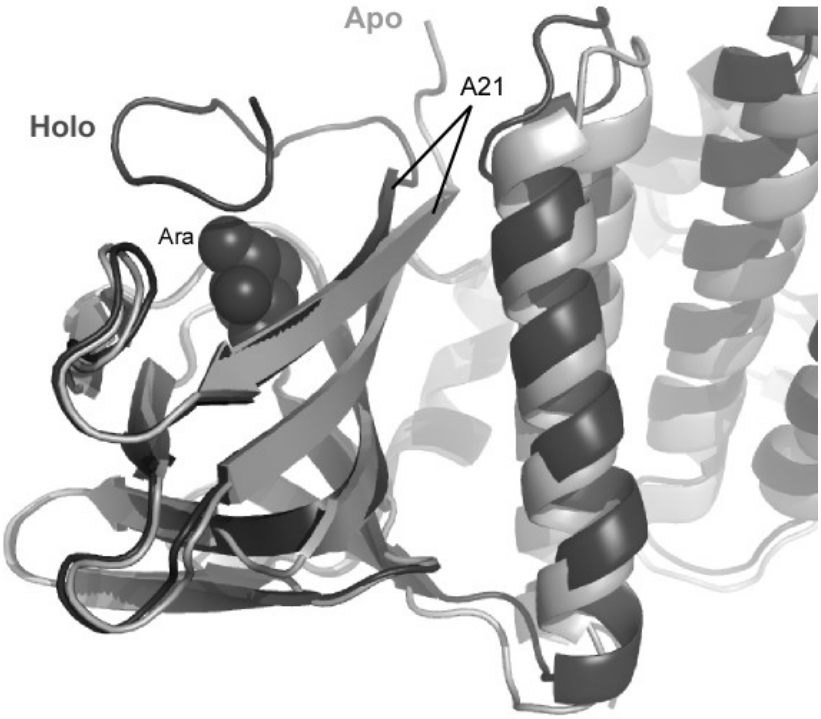

Figure 2. Overlay of apo and holo dimerization domain. Structures aligned by residues 23-122. Arabinose in holo structure shown in spheres.

Step two: Linker transitions out of helical state

In Chapter 3, I provided data consistent with the linker undergoing a helix-coil transition between the repressing (helical linker) and inducing (non-helical linker) forms of the protein. A preferentially helical linker is sufficient to bias AraC to the repressing state, while a non-helical linker does the opposite. Furthermore, AraC's short linker is incapable of forming a stable helix in isolation, but will readily continue a prenucleated helix.

An α-helical linker extending from the C-terminal helix of AraC's dimerization domain – as I propose exists in the repressing state – should rigidly hold the DBDs projected away from the dimerization domains. This structure allows for the DNA looping seen when AraC binds to distant sites to repress, but precludes binding to adjacent repeats, as would be required for induction. However, when the linker transitions into a flexible, non-helical state, binding to direct repeats is possible. This transition is key to the shift from repression to induction, as biasing the linker's structure to either state will thus favor the corresponding form of regulation (repression vs. induction) to the exclusion of the other.

The cascading rearrangement of the protein outlined above leads to the area immediately preceding the interdomain linker. As I have demonstrated, an interruption of the preceding helix will disrupt the helical structure of the linker, shifting it into a preferentially non-helical state. The changes propagated from the arm provide a clear avenue for such an interruption to occur. I therefore propose that this structural rearrangement reduces the helical propensity in the proximity of residue A166 (the base of the linker), in turn shifting the linker structure in preference of the flexible, non-helical conformation. This allows the DBDs to move into new locations, capable of binding to adjacent DNA repeats.

Step three: Interdomain contacts stabilize new conformation

Chapter 4 revealed the existence of heretofore unknown contacts between the DBD and the dimerization domain. Though previous work strongly suggested that such interdomain contacts would exist(Frato, Schleif 2009, Cole, Schleif 2012), I was only able to uncover their identity through hints offered

by bioinformatic examination of residue covariance. Screening these suggested residues showed not only the location of interdomain contacts, but that these contacts were needed specifically for induction.

Specifically, residue N194 appears to contact both H126 and N139 during induction. The locations of these residues readily allow for the construction of a plausible model with them in close proximity (Figure 3). This model brings N194 on the DBD in close proximity to H126 and N139 on opposite dimerization domain subunits. Furthermore, this model maintains the appropriate distance to facilitate the interdomain linker between the C terminus of the dimerization domain and the N terminus of the DBD, and also leaves the DNA binding helices of the DBD free to bind. Additionally, when the predicted contact surface is viewed as space-filling, a well-contoured interface is apparent, as shown in Figure 4. Such an interface would be energetically favorable, and is only possible when the linker is sufficiently flexible to accommodate it, preventing AraC from adopting such a conformation when in the repressive form, with its helical linker. Thus, a flexible, non-helical linker should automatically favor this structure, stabilizing the inducing state.

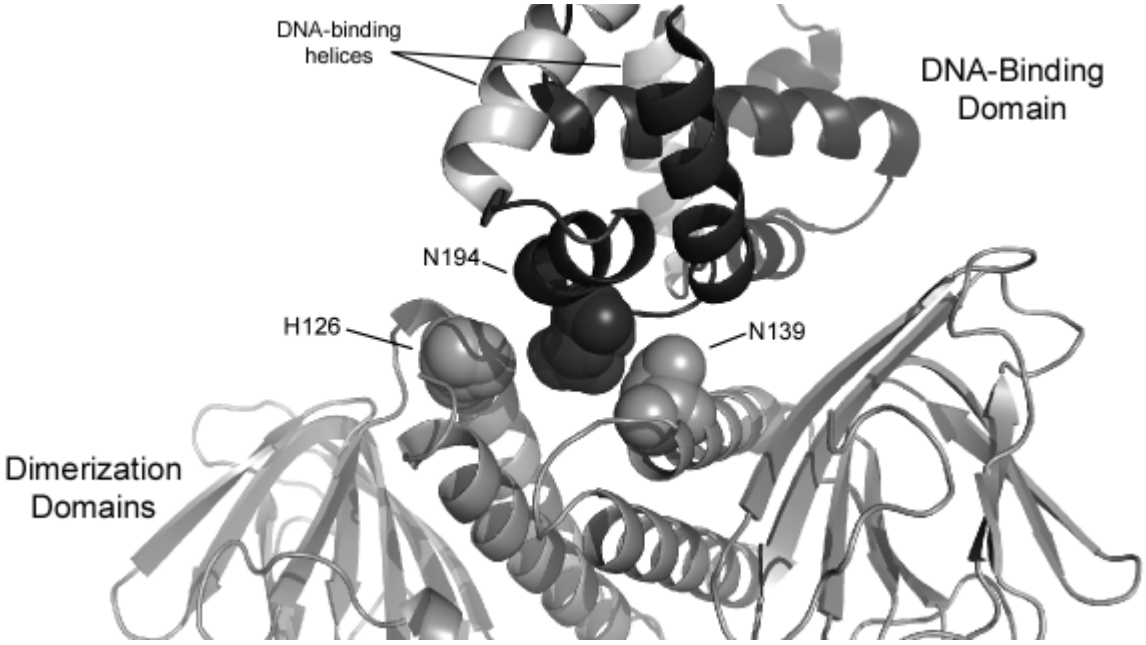

Figure 3. DBD-DD interdomain contacts

Notably, this model only accounts for one of the protein's two DBDs. However, as can be seen in Figure 1, if this contacting DBD binds to the I_1 site, this arrangement would position the other DBD in close proximity to I_2. This stabilized protein conformation functionally increases the local concentration of DBD in relation to I_2, increasing the effective binding affinity. This is especially noteworthy as the I_2 DNA sequence displays a 100-fold lower affinity for DBD than I_1. Such relative affinities could potentially offer an effective "failsafe," making the I_1-I_2-bound species only significantly populated when the protein is bound to arabinose.

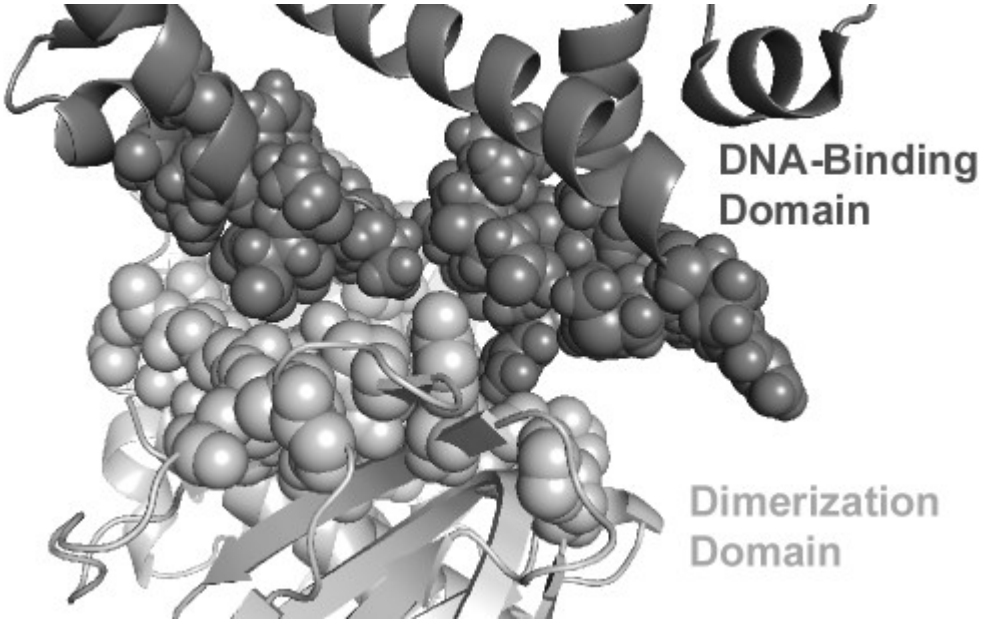

Figure 4. DNA-binding domain/dimerization domain interface.

References

COLE, S.D. and SCHLEIF, R., 2012. A new and unexpected domain-domain interaction in the AraC protein. *Proteins: Structure, Function, and Bioinformatics,* **80**(5), pp. 1465-1475.

DIRLA, S., CHIEN, J.Y. and SCHLEIF, R., 2009. Constitutive mutations in the *Escherichia coli* AraC protein. *Journal of Bacteriology,* **191**(8), pp. 2668-2674.

FRATO, K.E. and SCHLEIF, R.F., 2009. A DNA-assisted binding assay for weak protein-protein interactions. *Journal of Molecular Biology,* **394**(5), pp. 805-814.

ROSS, J.J., GRYCZYNSKI, U. and SCHLEIF, R., 2003. Mutational analysis of residue roles in AraC function. *Journal of Molecular Biology,* **328**(1), pp. 85-93.

SOISSON, S.M., MACDOUGALL-SHACKLETON, B., SCHLEIF, R. and WOLBERGER, C., 1997. The 1.6 A crystal structure of the AraC sugar-binding and dimerization domain complexed with D-fucose. *Journal of Molecular Biology,* **273**(1), pp. 226-237.

WELDON, J.E., RODGERS, M.E., LARKIN, C. and SCHLEIF, R.F., 2007. Structure and properties of a truely apo form of AraC dimerization domain. *Proteins,* **66**(3), pp. 646-654.

Appendix I: Python scripts

MSA-1line.py

```
#Takes a multiple sequence alignment file and outputs the same alignment each converting each sequence into a single line - for use as input for ResRand.py (next script)

#Input and output files located in “GREMLIN” folder by default

#Output written to “1line.txt”

#Run as "python MSA-1line.py [input file]"


import sys

myfile=open(sys.argv[1],"r")

myfilelist=myfile.readlines()

f1=open(r"GREMLIN\1line.txt","w+")


for l in myfilelist:
    if l[0]==">":
        print l
        f1.writelines(["\n%s" % l])
    else:
        f1.writelines([l[0:-1]])
```

ResRand.py

```
#Takes single-line alignment generated by MSA-1line.py and fills any gap characters with a random residue to generate file - for use as GREMLIN input

#Only works properly if input alignment is formatted into 1 line per sequence

#Input and output files located in “GREMLIN” folder by default

#Output written to “GapFill.txt”

#Run as "python ResRand.py [input file] [# of last arm residue]"
```

```
import sys
import random
myfile=open(sys.argv[1],"r")
myfilelist=myfile.readlines()

ArmLimit=int(sys.argv[2])
#This line is unnecessary for final script, holdover from use in generating a list of arm residue contacts

f=open(r"GREMLIN\1line.txt","w+")




for l in myfilelist:
    if l[0]==">":
        f.writelines(["\n%s" % l])
    else:
        f.writelines([l[0:-1]])

f1=open(r"GREMLIN\1line.txt","r")
f2=open("GREMLIN\GapFill.txt","w+")
linefile=f1.readlines()

ResList=["A","C","D","E","F","G","H","I","K","L","M","N","P","Q","R","S","T","V","W","Y"]
#Creates list of 20 possible residues to substitute for gap character

newres=random.randint(0,19)
#Selects a randomly-generated residue from above list to be substituted

x=0
for l in linefile:
    if l[0]==">":
```

```
        f2.writelines(["\n%s" % l])
        print l
    else:
        for i in l:
            if str(i)=="-":
                f2.write(ResList[random.randint(0,19)])
            else:
                f2.write(i)

        x=x+1
#Replaces all instances of gap characters (“-“) with randomly-selected residue from above list

print(str(x) + ' sequences')
#Reports number of sequences in alignment
```

AllContacts.py

#Takes output from GREMLIN and formats into lists of all covarying pairs with reported residue numbers replaced with appropriate residue numbers for E. coli AraC, as well as reported scores - for import into Excel spreadsheet

#Each line represents a single pair and associated data

#Input and output files located in “GREMLIN” folder by default

#Output written to “ContactsALL.txt”

#Run as "python AllContacts.py [input file]"

```
import math
import sys

ResDict={91:1,92:2,93:3,94:4,95:5,96:6,97:7,98:8,99:9,100:10,102:11,103:12,104:13,107:14,108:15,109:16
,110:17,111:18,112:19,113:20,114:21,115:22,116:23,117:24,118:25,119:26,120:27,121:28,122:29,129:30,1
30:31,131:32,132:33,133:34,134:35,135:36,136:37,137:38,138:39,139:40,140:41,141:42,143:43,144:44,145
:45,146:46,147:47,148:48,149:49,150:50,151:51,152:52,153:53,154:54,155:55,156:56,157:57,158:58,159:5
9,163:60,164:61,165:62,167:63,168:64,169:65,170:66,171:67,172:68,173:69,174:70,175:71,176:72,177:73,
178:74,179:75,180:76,181:77,182:78,183:79,184:80,185:81,186:82,187:83,188:84,189:85,194:86,195:87,19
```

```
6:88,197:89,198:90,199:91,200:92,201:93,202:94,203:95,204:96,205:97,206:98,207:99,208:100,209:101,21
0:102,211:103,212:104,213:105,214:106,215:107,216:108,217:109,218:110,219:111,220:112,221:113,222:1
14,228:115,248:116,249:117,250:118,251:119,252:120,253:121,254:122,255:123,256:124,257:125,262:126,
263:127,264:128,265:129,266:130,267:131,268:132,269:133,270:134,271:135,272:136,273:137,274:138,27
5:139,276:140,293:141,294:142,295:143,298:144,299:145,300:146,301:147,302:148,303:149,304:150,305:1
51,306:152,307:153,308:154,309:155,310:156,311:157,312:158,313:159,314:160,316:161,317:162,318:163,
319:164,320:165,321:166,322:167,323:168,324:169,325:170,331:171,332:172,333:173,334:174,335:175,33
6:176,337:177,338:178,339:179,340:180,341:181,342:182,343:183,344:184,345:185,346:186,347:187,348:1
88,349:189,354:190,355:191,356:192,357:193,358:194,359:195,360:196,361:197,362:198,363:199,364:200,
365:201,366:202,373:203,374:204,375:205,376:206,377:207,378:208,379:209,380:210,381:211,382:212,38
3:213,384:214,385:215,386:216,387:217,388:218,389:219,390:220,391:221,392:222,393:223,394:224,395:2
25,396:226,397:227,398:228,399:229,400:230,401:231,402:232,403:233,404:234,405:235,406:236,407:237,
408:238,409:239,410:240,411:241,412:242,413:243,414:244,415:245,416:246,417:247,418:248,419:249,42
0:250,421:251,422:252,423:253,424:254,425:255,426:256,427:257,428:258,429:259,430:260,431:261,432:2
62,433:263,434:264,435:265,436:266,437:267,438:268,439:269,440:270,441:271,442:272,443:273,444:274,
445:275,446:276,447:277,448:278,449:279,450:280,451:281,452:282,453:283,454:284,455:285,456:286,45
7:287,458:288,459:289,460:290,461:291,462:292}

#Generates dictionary to translate an alignment position into the appropriate residue number for E. coli AraC



myfile=open(sys.argv[1],"r")

myfilelist=myfile.readlines()

f=open("GREMLIN\contactsALL.txt","w+")

f.write("i\tj\tj(2k9s)\ts_sco\tprob\n")

#Creates column header labels (tab-delimited)


x=0

for l in myfilelist:

    linelist=l.split()

    if int(linelist[0])>1:

        if int(linelist[0]) in ResDict:

            if int(linelist[1]) in ResDict:

                if int(linelist[1])>1:

                    f.writelines(["%s\t" % ResDict[int(linelist[0])]])

                    f.writelines(["%s\t" % ResDict[int(linelist[1])]])

                    f.writelines(["%s\t" % int(ResDict[int(linelist[1])]-174)])

                    f.writelines(["%s\t" % linelist[5]])

                    f.writelines(["%s\n" % linelist[6]])
```

```
            x=x+1
#Generates list of first residue in pair, second residue in pair, residue number in the DBD structure for second residue (negative if DD), reported S score, and reported probability (tab-delimited)


f.close()
print("written to contactsALL.txt")
print(x)
```

Appendix II: Arabinose Alters Both Local and Distal H-D Exchange Rates in the *Escherichia coli* AraC Transcriptional Regulator

Important Note: The large majority of the work presented in this appendix was performed by Alexander Tischer in the laboratory of Matthew Auton at the Mayo Clinic, based on my general experimental designs, and written primarily by Matthew Auton. This work was submitted to *Biochemistry* as an accompanying manuscript to Brown & Schleif 2019 (represented in Chapter 3), and published alongside it. The manuscript is presented here as-is.

Arabinose Alters Both Local and Distal H-D Exchange Rates in the *Escherichia coli* AraC Transcriptional Regulator

Alexander Tischer[1], Matthew J. Brown[2], Robert F. Schleif[2]*, and Matthew Auton[1]* March 14, 2019

[1]Division of Hematology, Departments of Internal Medicine and Biochemistry and Molecular Biology, Mayo Clinic, Rochester, Minnesota, 55905, USA.

[2]Department of Biology, Johns Hopkins University, Baltimore, Maryland , 21218, USA.

* To whom correspondence is addressed, auton.matthew@mayo.edu and rfschleif@jhu.edu.

Author Contributions: AT performed and analyzed the Hydrogen-Deuterium Exchange Mass Spectrometry (HXMS) experiments. MJB expressed and purified the AraC dimerization domain and the full-length AraC protein. RFS and MA designed the research. MA wrote the paper.

Running Title: H-D Exchange in AraC.

Keywords: HXMS, Allostery, Arabinose, Operon, AraC, Dimerization Domain, DNA Binding Domain, Linker, Helix Capping.

Abbreviations Used: HXMS, Hydrogen Deuterium Exchange Mass Spectrometry; ...

Abstract

In the absence of arabinose, the dimeric *Escherichia coli* regulatory protein of the L-arabinose operon, AraC, represses expression by looping the DNA between distant half-sites, $araO_2$ and $araI_1$. Arabinose binding to the dimerization domains transition AraC to preferentially bind two adjacent DNA half-sites, $araI_1$ and $araI_2$, where it stimulates RNA polymerase transcription of the *araBAD* catabolism genes. Genetic and biochemical experiments have hypothesized that arabinose allosterically induces a helix-coil transition of a short linker between the dimerization and DNA binding domains which enables the conformational switch to the inducing state (Brown & Schleif 2019, accompanying manuscript) [1]. Hydrogen-deuterium exchange mass spectrometry (HXMS) was utilized to test this hypothesis and determine the structural regions involved in the conformational activation of AraC by arabinose. Comparison of the HXMS data of the individual dimeric dimerization domains and the full-length dimeric AraC protein in the presence and absence of arabinose do not support the a helix-coil transition nor the presence of any substantial local unfolding of protein secondary structures that could control such a transition. Rather, HXMS data show a prominent arabinose-induced destabilization of the amide hydrogen bonded structure of linker residues (I_{167}-N_{168}) that could be the result of an increased probability of forming a helix capping motif at the C-terminal end of the major dimerizing helix of the dimerization domain that immediately preceeds the interdomain linker. These conformational changes could allow for quaternary repositioning of the DNA binding domains required for induction of the *araBAD* promoter through rotation of peptide backbone dihedral angles of just a couple residues. Additionally, subtle changes in exchange rates are visible around the arabinose binding pocket and in the DNA binding domain.

Introduction

The homodimeric *Escherichia coli* protein, AraC [2–4], regulates transcription of the *araBAD* genes required for the catabolism of L-arabinose as well as the *araC* gene itself [5]. The *araBAD* genes and the *araC* gene are transcribed in opposite directions yielding mRNAs that are translated into the enzymes required for arabinose catabolism and the AraC gene regulator, **Fig.1A** [6]. AraC consists of a dimerization domain that binds arabinose [7,8] and a DNA binding domain [9] connected by an interdomain linker [10,11]. Regulation of the arabinose operon is achieved by the binding of arabinose to AraC which facilitates its switching from a repressor to an activator of transcription involving five DNA binding half-sites, the adjacent I_1 and I_2 half-sites of *araI*, 35 base pairs upstream of the p_{BAD} transcription start site, two half-sites at $araO_1$ 120 base pairs upstream of p_{BAD}, and the half-site at $araO_2$ that is 245 base pairs upstream of p_{BAD} [12].

In the absence of arabinose, AraC predominantly loops the DNA between I_1 and O_2 and represses the transcription of both the genes encoding itself and the catabolic enzymes [13]. AraC can also autoregulate its own synthesis in the presence of arabinose by binding to the adjacent half-sites of O_1 which overlap the p_C RNA polymerase binding region [14–16]. The repressed state involving DNA looping is dependent on supercoiling of the bacterial DNA [17, 18]. The binding of arabinose to the dimerization domains causes AraC to switch from a DNA looped O_2-I_1 bound configuration to an unlooped configuration in which AraC binds the adjacent *araI* sites, I_1 and I_2, **Fig.1A** [19]. This unlooped DNA state stimulates RNA polymerase to bind and initiate transcription from p_{BAD}.

The shift from repressing to inducing requires a substantial quaternary structural rearrangement of the domains of AraC, **Fig.1A** [19]. Many of the recent studies of AraC have been directed at understanding the mechanisms by which the binding of arabinose brings about these structural changes [20–22]. The binding of arabinose to the protein's dimerization domains has been proposed to reduce the rigidity with which the DNA-binding domains are held in positions and orientations favoring DNA looping. A more flexible state of the protein reduces the energetic cost of repositioning the DNA binding domains required for binding to adjacent DNA half-sites, and much of the repressed, DNA looped, species is then replaced by DNA with

AraC bound at $I_1 - I_2$ [16]. Two regions of AraC have been implicated to be involved in such a structural rearrangement.

1) Crystallographic structures of the dimerization domain with and without bound arabinose show the N-terminal arm (first 20 residues of AraC) in two different conformations [8, 23]. The binding of arabinose repositions the arm so that it binds directly over the bound arabinose. This conformational change has been proposed to be the first step of a mechanism that changes the protein's flexibility. Consistent with this notion is the fact that almost all of the mutations that reduce repression by AraC lie in the arms [24, 25]. Deletions also reduce or eliminate repression [22]. These facts imply that in the absence of arabinose, the arm plays a key role in holding the protein in its less flexible, repressing state, and that the arm's repositioning in the presence of arabinose ends this activity [20].

2) The second region that appears to be involved in the repressing-inducing conformation shift of AraC is the eight residue linker [11] between the dimerization domain and the DNA binding domain. This linker is malleable with respect to mutations it can accommodate [10], but recent studies suggest it plays an active role in AraC's conformational switch. Mutations have been identified in this region that decrease repression, increase induction of p_{BAD}, and possibly alter the DNA half-site binding preference [26]. Fluorescence anisotropy measurements have also shown that arabinose binding to AraC lacking the DNA binding domains increases the flexibility of the linker [27]. Furthermore, recent studies substituting poly-alanine or poly-glycine into the linker and insertions and deletions in the linker indicate that in the absence of arabinose, the linker may be helical, and lead to the hypothesis that the addition of arabinose destabilizes the helix [1], a hypothesis which motivated the current studies.

Here, we query the arabinose free and bound states of AraC in the absence of DNA using hydrogen-deuterium exchange mass spectrometry (HXMS) on the assumption that the properties of AraC relevant to its conformational shift in the presence of arabinose, (**Fig.1B**), can be observed in the absence of DNA. HXMS [28] measures the steady-state time-dependent accumulation of deuterium into the peptide backbone. When diluted into ^2H_2O buffer, protein amide hydrogens exchange with solvent deuterium according to the local conformational dynamics of the peptide backbone. Structurally protected amide hydrogens in stable H-

bonded secondary structure (α-helix or β-sheet) exchange slowly, while amides in flexible regions of structure exchange quickly.

Arabinose binding predominantly increases the amide exchange rates of residues Ile_{167} and Asn_{168} following the C-terminal α2-helix of the dimerization domain near the beginning of the interdomain linker. Contrary to the proposed order to disorder helix-coil transition of this linker [1], the observed changes instead support a mechanism whereby arabinose binding destabilizes the junction between the dimerizing helix and interdomain linker possibly through the formation of a C-terminal helix cap. Additionally, arabinose stabilizes both the dimerization domains in regions adjacent to the arabinose binding site and the DNA binding domains in the central α4-helices demonstrating that the DNA binding domains are indeed responsive to the allosteric effect of arabinose on the interdomain linker even in the absence of DNA.

Methods

Protein Expression and Purification.

Full-length AraC and dimerization domain was overproduced by the pET24 expression vector(Novagen), isolating the protein by sequential heparin and HiTrap-Q ion exchange columns as previously described [29]. AraC dimerization domain was overproduced by the pET21 expression vector using the first 182 residues of the AraC protein plus a C-terminal His-6 tag, isolated using a Ni-NTA column and trypsin digestion, followed by a HiTrap-Q column as previously described [23].

Hydrogen Deuterium Exchange Mass Spectrometry (HXMS).

HXMS was performed on a Thermo Scientific LTQ Orbitrap XL mass spectrometer coupled to a Waters Nano Acquity UPLC system. HX samples were prepared by dilution of 100 μL of protein solution (~20 μM AraC dimerization domain in Tris buffered saline, pH 7.4 with or without 20 mM L-arabinose) into 400 μL D_2O buffer (TBS, pD 7.4 with or without 20 mM L-arabinose). Samples were mixed and the exchange was quenched by a drop in pH to 2.7 at defined time points ranging from 10 s to overnight incubation.

Quenched samples were mixed again and injected into a 200 μL loop within the cooling box containing the LC setup at ~0 ˚C. After manual injection the protein was passed through a custom packed pepsin column (250 μL) [30] for non-specific digestion at a flow rate of 200 μL/min using 0.1% formic acid (pH 2.7). Peptides were collected on a Higgins Analytical Targa C8 trap column. After 5 min a second valve was used to switch the trap column from isocratic flow in line with a water/acetonitrile gradient (2 μL/min, 1-40% acetonitrile in 18min, followed by a steep gradient up to 90% in 2 min). The gradient elutes peptides from the trap column while a subsequent Waters Symmetry C18 column separates them prior to injection into the electrospray emitter of the mass spectrometer. The mass spectrometer was operated in positive polarity mode at a resolution of 60,000.

Prior to the measurement of HD exchange kinetics, an initial peptide mapping of AraC dimerization domain and full-length AraC (see Supporting Information **SI-Fig.1**) were performed using protein dialyzed into 0.1% formic acid pH 2.7. Multiple ”all H” runs were performed using exclusion lists containing well defined and strong peptide peaks. This procedure provides a more complete peptide map of the protein as it allows the determination of weaker, less populated peptides. Peptides were identified using Bioworks 3.3.1 (Thermo Fisher Scientific) using a search tolerance of 4 ppm. Exclusion lists were generated with EXMS2 [31] using peptides with a PPep score of ≥0.5. The final peptide map was generated by combining all successfully identified peptides using a PPep score of ≥0.99.

EXMS2 [31] was used to identify deuterated peptides using the peptide map. The m/z tolerance was 10 ppm, the retention time window 4 min for the ”all H” sample and 2 min for deuterated samples. The individual and the summed peak noise thresholds were set to 500 and 1500 respectively. HDsite [32] was performed using an experimental temperature of 25 ˚C and a pD of 7.4. The deuteration range was set to 0 to 0.8. Switchable peptides were averaged manually after analysis.

Results.

Hydrogen-deuterium exchange of the dimerization domain.

HXMS was performed on the AraC dimerization domain at 20 μM concentration in the presence and absence of 20mM arabinose (1000-fold excess) to compare the dynamics of AraC in its apo and arabinose bound states. The affinity of AraC for arabinose has been previously determined to be K_D = 0.3±0.1 μM [26, 29]. In the current study, the dimerization domain in its apo and bound states exists as a dimer as at even at picoMolar AraC concentrations, the protein remains dimeric [33].

The structure of the AraC dimerization domain (**Fig.2A**) consists of ten β-strands that loop around to make up the arabinose binding barrel [8]. This binding barrel is flanked by a 3_{10} helix and two consecutive anti-parallel C-terminal α-helices (residues E_{124}-I_{167}) that form the dimerization interface. The first residue of the interdomain linker connecting the dimerization domain to the DNA binding domain immediately follows the second helix, α2 [26].

Fig.2B illustrates the exchange fraction of deuterium incorporated into the arabinose bound and free states of the dimerization domain at three D_2O incubation times; 1 min, 1 hr, and 24 hr as a function of residue number. Additional time points ranging between 10 s and 24 hr are presented in Supporting Information **SI-Fig.2**. HX occurs throughout the solvent-exposed secondary structures to different extents depending on the local peptide backbone dynamics. The fraction deuterium incorporated into the domain is greatest in the N-terminus, β2 loop (residues I_{36}-I_{40}), β4-β5 hairpin (I_{51}-F_{64}), β8 loop (Y_{97}-P_{100}), 3_{10} helix (A_{102}-L_{108}), and the C terminus. HX within the α1-helix (residues E_{124}-Q_{142}) is minimal and residues in the α2-helix (A_{152}-R_{162}) do not exchange at all indicating that these helices are stable. The time dependence of deuterium incorporation into specific residues in the various secondary structures regions of AraC is presented in Supporting Information **SI-Figs.3 & 4**.

Here, we focus on the structural regions whose dynamics respond to the binding of arabinose. **Fig.2** shows that HX of much of the domain remains the same in both the apo and arabinose bound states, but in three structural regions of the domain, the kinetics of exchange were clearly different in the presence and absence of arabinose. These regions include two sides of the arabinose binding barrel, 1) the β4-β5 hairpin (I_{51}-F_{64}) and 2) the loop between the β8 strand and the 3_{10} helix adjacent in tertiary structure to Ala_{17} in the N-terminal arm. The C-terminal interdomain linker, partially resolved in various crystal structures of the

dimerization domain, is also responsive to arabinose binding.

Simultaneous stabilization and relaxation of the arabinose binding barrel.

Figs.3 & 4 illustrate the dynamics of exchange in the arabinose binding barrel in greater detail. In **Fig.3**, the β4-β5 hairpin loop (residues I_{51}-F_{64}) is observed to exchange less in the bound state relative to the apo state indicating that arabinose binding stabilizes this secondary structure region. Deuterium incorporation is observed to increase as a function of time in the apo state within specific peptides (T_{50}-E_{63}, G_{53}-E_{63}, and

G_{55}-E_{63}) of this structural region causing a gradual shift in relative mass. Conversely, the mass of these peptides remains comparably constant in the bound state, **Fig.3A**. Residues (I_{51}, G_{55}-V_{56}, and Q_{60}-E_{63}) in this structural region show the same gradual uptake of deuterium in the apo state relative to the bound state as a function of time (**Fig.3B**). Overall, exchange of residues in the entire β4-β5 hairpin loop increases faster in the apo state (**Fig.3C**) indicating that this loop readily samples open and closed conformations that, possibly, facilitates the binding of arabinose **Fig.3D**. Once arabinose is bound, the closed conformation of this hairpin loop is stabilized and HX of the region is quenched.

In addition to the stabilization of the β4-β5 hairpin loop caused by the binding of arabinose, residues on the opposite side of the binding barrel in the β8-3_{10} loop and in A_{17} of the N-terminus **Fig.4B**, each of which are adjacent in the three dimensional structure of the domain, show increased exchange kinetics. **Fig.4A** shows that the relative masses of two peptides (F_{98}-E_{106} and V_{96}-W_{107}) in the β8-3_{10} loop and one peptide (F_{15}-L_{19}) in the N-terminal arm are slightly shifted to higher masses as a function of time in the bound state. Closer examination of the kinetics of residues Y_{97}-R_{99} and A_{17}, shows that the exchange is shifted to earlier times, indicating arabinose binding relaxes this region of the binding barrel resulting enhanced exchange rates, **Fig.4C & D**. Thus, the flexibility of this structural region is increased concomitant with stabilization of the β4-β5 hairpin loop. None of the residue sidechains in either β4-β5 hairpin loop, the β8-3_{10} loop, or A_{17} make direct atomic interactions with arabinose and thus could also be considered allosteric effects within the binding barrel.

<u>Allosteric relaxation of the C-terminal interdomain linker.</u>

In addition to changes in the exchange rates of residues lining the arabinose binding barrel, we also see alterations in exchange rates of more distant residues. **Fig.5** illustrates that residues in the interdomain linker also relax in response to arabinose binding. This is observed in a specific peptide spanning residues R_{161}-S_{169} which incorporates deuterium faster in the bound state than in the apo state, **Fig.5A**. The exchange kinetics of residues are increased resulting in a shift in the exchange fraction versus time to earlier time points, **Fig.5B**. This allosteric effect is highly localized to I_{167}-N_{168} as the surrounding residues exchange equally in

both the bound and apo states of the domain, **Fig.5C**.

Upon close inspection of hydrogen bonded structure of the the C-terminal helix (B chain of pdbID = 2ARC, arabinose bound state) [8], it becomes evident that the helix is capped at the C-terminus by a hydrogen bond between the N_{168} side chain and the M_{164} carbonyl, **Fig.5D**. Following the nomenclature of Aurora & Rose [34], residues prior to and including I_{167} (the C-cap) follow the ($i \longrightarrow i + 4$) traditional hydrogen bonding pattern of an α-helix and N_{168} is the first nonhelical residue, C'. Capping of carbonyls by side chains is less frequent than by backbone amides at helix C-termini [34], but this cap redistributes the α-helix ($i \longrightarrow i + 4$) hydrogen bonding pattern to ($i \longrightarrow i + 3$) hydrogen bonds between the carbonyl of A_{166} and the amide of E_{169} and between the carbonyl of E_{165} and the amide of N_{168} where the helix terminates. Apo structures of AraC either do not have N_{168} resolved or N_{168} does not form a C-cap motif [7, 23]. This suggests that formation of the C-cap may be responsible for the enhanced exchange of residues I_{167}-N_{168} in the presence of arabinose. C-capping specifically by asparagine has also been observed in other proteins using NMR [35].

Hydrogen-deuterium exchange of full-length AraC confirms observations in the dimerization domain and reveals allosteric stabilization of the DNA binding domain.

Fig.6A shows the exchange fraction of deuterium incorporated into the full-length AraC protein at three D_2O incubation times; 10 s, 1 min, and 10 min as a function of residue number. Additional time points ranging between 10s and 1hr are presented in Supporting Information **SI-Fig.5**. The kinetics of deuterium incorporation into the individual residues of different secondary structures regions of full-length AraC are presented in Supporting Information **SI-Figs.6 & 7**. Longer D_2O incubation times resulted in the aggregation of AraC that caused loss of peptide coverage and resolution of the protein hydrogen exchange. This experimental complication was only marginally reduced by the presence of bound arabinose and 20% glycerol [36] and is likely due to a kosmotropic effect of D_2O compared to water as the hydrogen bonding of D_2O is weaker than H_2O [37]. Hydrogen exchange at shorter incubation times was amenable to data analysis and interpretation.

Fig.6A illustrates that HX within the DNA binding domain is significantly greater than that of the dimerization domain. The NMR solution structure of the AraC DNA binding domain is comprised of seven α- helices [38]. The enhanced HX the DNA binding domain indicates that it is less stable than the dimerization domain, certainly more dynamic, or even partially disordered since the α-helices are exchanging very rapidly throughout the DNA binding domain structure, **Fig.6B**.

Although the stabilization of the β4-β5 hairpin loop by arabinose binding was not observed in the full-length protein in the short time range of 10 s to 30 min, the relaxation of the arabinose binding barrel and the interdomain linker were observed. **Fig.6C** illustrates these observations with two peptides; F_{98}-E_{106} in the β8-3_{10} loop and R_{161}-S_{169} in the interdomain linker. The difference between relative mass shifts of these peptides in the presence and absence of arabinose as a function of time is subtle, but the exchange of individual residues reveals enhanced uptake of deuterium in the Y_{97}-R_{99} residues of the β8-3_{10} loop and a kinetic shift of deuterium incorportation in residues R_{163}-N_{168} of the interdomain linker. These observations confirm that the intrinsic conformational dynamics of the dimerization domain alone also apply to the full-length AraC protein containing the DNA binding domain.

Despite the high exchange rates seen in the DNA binding domain, it was possible to see a significant reduction in the exchange rates of residues in the α4-helix caused by the presence of arabinose. The relative mass shifts of residues S_{225}-L_{237} in the top panel of **Fig.6D** demonstrate that deuterium incorporation into the α4-helix is faster in the absence of arabinose. Site-resolved regions A_{235}-L_{237} and R_{227}-Q_{234} also show a kinetic delay of the exchange reaction in the presence of arabinose. Comparatively, residues S_{262}-F_{276} of the C-terminal peptide Y_{260}-F_{276} derived from helices α6-7 are highly dynamic with no observed change induced by arabinose.

Discussion.

Here, we have described hydrogen-deuterium exchange experiments designed to identify which structural regions of AraC respond to arabinose binding. HXMS was performed as a function of time between 10 s

and 24 hrs to assess the kinetics of exchange in the dimerization domain with the interdomain linker and full-length AraC, containing both the dimerization and the DNA binding domains. Two features are notable in this comparative study of AraC in the presence and absence of excess arabinose. 1) HX differences are observed in residues of the arabinose binding dimerization domain that confirm arabinose is bound. 2) As expected for an allosteric protein, arabinose induced changes in HX dynamics are observed both adjacent to the bound arabinose and at more distant locations in the protein.

Arabinose binding directly affects the exchange dynamics of residues surrounding the arabinose binding pocket. HXMS detects a quenching of the hydrogen deuterium exchange in residues of the β4-β5 hairpin loop (I_{51}-F_{64}) near the top of the arabinose binding β-barrel, (**Fig.3**), indicating a stabilization of this region of the dimerization domain when arabinose is bound. A moderate enhancement of the exchange kinetics is observed in residue A_{17} of the N-terminal arm and in residues V_{96}-W_{107} of the β8-3_{10} loop region, (**Fig.4**). This may reflect a repositioning of the N-terminal arm of the dimerization domain as is observed in crystal structures [21, 23]. However, other than that observed for A_{17}, there is no detectable difference in the exchange of residues of the N-terminal arm in the time range measured. Exchange in residues 4-12 is complete by 10s and residues 13-16 are stable and do not exchange regardless of whether arabinose is bound (**SI-Fig.3**). Moreover, the N-terminus is unaffected by arabinose in the full-length AraC containing the DNA binding domain (**SI-Fig.6**).

The physiological and biochemical behavior of poly-alanine and poly-glycine linker substitutions [1] and proline mutations at different positions of the interdomain linker as well as insertions and deletions in the linker [26, 27] have suggested that, in the absence of arabinose, the linker is helical and hypothesize that arabinose binding induces a helix-coil transition that provides the flexibility necessary for the protein to tightly bind to adjacent half-sites and activate transcription. The HXMS experiments reported here do not provide evidence supporting the helix-coil hypothesis because the extent of exchange within linker residues E_{169}-H_{172} is not substantially different between the arabinose bound and unbound states. However, arabinose binding does result in highly localized differences in exchange kinetics at the beginning of the interdomain linker starting at the C-terminal end of the α2-helix of the dimerization domain.

As mentioned above, the exchange rates of residues I_{167}-N_{168} are significantly increased when arabinose is bound, (**Fig.5**). One mechanism that might account for this is an arabinose-induced formation of a C-terminal helix cap by the asparagine resulting from weakened amide hydrogen bonds at the end of the α2-helix. Similar enhanced HX has been previously observed by NMR in a destabilized helix N-capping box of helix G of the C-lobe of cardiac troponin C upon binding the phosphomimetic N-domain of troponin I [39]. A C-terminal helix capping motif terminating the α2-helix of the AraC dimerization domain in the presence of arabinose could increase the freedom of several of the peptide backbone dihedral angles of residues following the N_{168} C-cap motif and allow AraC to more easily adopt the inducing conformation. Several facts argue in favor of this interpretation. First, asparagine is one of the strongest C-terminal capping residues [34, 35]. Second, in the crystal structures of AraC dimerization domain with and without bound arabinose, C-terminal capping only occurs in the protein crystallized in the presence of arabinose.

It should be noted that a somewhat elevated rate of exchange was also observed at residues I_{167}-N_{168} even in the absence of arabinose. It is possible, therefore, that the sidechain of N_{168} is in an equilibrium between a capped state and an uncapped state so that exchange can occur in the absence of arabinose, and that the binding of arabinose raises the probability of the capped state. This flickering between the normal helix hydrogen bonding and the sidechain capping the of dimerizing helix might allow for a modulation of the length of the α2-helix or the linker region just beyond a few residues and serve to regulate the strength by which the DNA binding domains are held in positions favorable for DNA looping and repression.

In these experiments, the binding of arabinose to the dimerization domain also decreases the exchange rate of multiple residues in the DNA binding domain, particularly in the domain's central alpha helix that supports the two helix-turn-helix DNA binding motifs. This must result from an altered DNA binding domain-dimerization interaction. Whether this is a result of a domain-domain interaction that only occurs in the presence or only in the absence of arabinose, or whether arabinose changes interactions that occur in both states cannot be determined with the current data. It should be noted however, that multiple lines of physical evidence exist for an interaction that occurs in the presence of arabinose [40]. Similarly, covarying amino acid residues in the dimerization domain and DNA binding domain in AraC homologs (M. Brown,

Ph.D. thesis [41]) occur only in positions compatible with the positioning of the DNA binding domains that is required for induction.

In conclusion, these HXMS studies of the WT sequence of AraC demonstrate that the binding of arabinose induces highly localized changes in H-D exchange rates allosteric to the binding site. These effects do not appear to involve any substantial unfolding of the protein, but are consistent with the formation of a C-terminal helix cap of the dimerization domain α2-helix. Although the mechanism of signal transmission to the responsive residues that arabinose has bound remains undetermined, such an allosteric effect on the backbone hydrogen bonded stability of residues at the junction between the α2-helix and the interdo- main linker would allow the DNA binding domains to unhinge and accommodate alternative DNA binding configurations commensurate with transcription activation of *araBAD* [12, 15, 42].

Acknowledgements

We thank Dr. George Rose for encouragement and ongoing discussions over the years. We also thank Drs. S. Walter Englander and Leland Mayne for very helpful scientific discussions regarding optimization of HXMS and the establishment of HXMS technology in the Auton lab.

References

[1] M. Brown and R. Schleif, "Accompanying manuscript," *Biochemistry*, vol. This Issue, 2019.

[2] D. Steffen and R. Schleif, "In vitro construction of plasmids which result in overproduction of the protein product of the arac gene of escherichia coli.," *Mol Gen Genet*, vol. 157, no. 3, pp. 341–344, 1977.

[3] W. Hendrickson and R. Schleif, "A dimer of arac protein contacts three adjacent major groove regions of the arai dna site.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 82, no. 10, pp. 3129–3133, 1985.

[4] S. Bustos and R. Schleif, "Functional domains of the arac protein.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 90, no. 12, pp. 5638–5642, 1993.

[5] D. Sheppard and E. Englesberg, "Further evidence for positive control of the l-arabinose system by gene arac.," *J. Mol. Biol.*, vol. 25, no. 3, pp. 443–454, 1967.

[6] G. Wilcox, J. Boulter, and N. Lee, "Direction of transcription of the regulatory gene arac in escherichia coli b-r.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 71, no. 9, pp. 3635–3639, 1974.

[7] S. Soisson, B. MacDougall-Shackleton, R. Schleif, and C. Wolberger, "The 1.6 a crystal structure of the arac sugar-binding and dimerization domain complexed with d-fucose.," *J. Mol. Biol.*, vol. 273, no. 1, pp. 226–237, 1997.

[8] S. Soisson, B. MacDougall-Shackleton, R. Schleif, and C. Wolberger, "Structural basis for ligand-regulated oligomerization of arac.," *Science*, vol. 276, no. 5311, pp. 421–425, 1997.

[9] S. Cole and R. Schleif, "A new and unexpected domain-domain interaction in the arac protein.," *Proteins*, vol. 80, no. 5, pp. 1465–1475, 2012.

[10] R. Eustance and R. Schleif, "The linker region of arac protein.," *J. Bacteriol.*, vol. 178, no. 24, pp. 7025– 7030, 1996.

[11] R. Eustance, S. Bustos, and R. Schleif, "Reaching out. locating and lengthening the interdomain linker in arac protein.," *J. Mol. Biol.*, vol. 242, no. 4, pp. 330–338, 1994.

[12] L. Huo, K. Martin, and R. Schleif, "Alternative dna loops regulate the arabinose operon in escherichia coli.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 85, no. 15, pp. 5444–5448, 1988.

[13] T. Dunn, S. Hahn, S. Ogden, and R. Schleif, "An operator at -280 base pairs that is required for repression of arabad operon promoter: addition of dna helical turns between the operator and promoter cyclically hinders repression.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 81, no. 16, pp. 5017–5020, 1984.

[14] M. Casadaban, “Regulation of the regulatory gene for the arabinose pathway, arac.,” *J. Mol. Biol.*, vol. 104, no. 3, pp. 557–566, 1976.

[15] S. Hahn and R. Schleif, “In vivo regulation of the escherichia coli arac promoter.,” *J. Bacteriol.*, vol. 155, no. 2, pp. 593–600, 1983.

[16] T. Harmer, M. Wu, and R. Schleif, “The role of rigidity in dna looping-unlooping by arac.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 2, pp. 427–431, 2001.

[17] S. Hahn, W. Hendrickson, and R. Schleif, “Transcription of escherichia coli ara in vitro. the cyclic amp receptor protein requirement for pbad induction that depends on the presence and orientation of the arao2 site.,” *J. Mol. Biol.*, vol. 188, no. 3, pp. 355–367, 1986.

[18] J. Borowiec, L. Zhang, S. Sasse-Dwight, and J. Gralla, “Dna supercoiling promotes formation of a bent repression loop in lac dna.,” *J. Mol. Biol.*, vol. 196, no. 1, pp. 101–111, 1987.

[19] J. Carra and R. Schleif, “Variation of half-site organization and dna looping by arac protein.,” *EMBO J.*, vol. 12, no. 1, pp. 35–44, 1993.

[20] M. Rodgers, N. Holder, S. Dirla, and R. Schleif, “Functional modes of the regulatory arm of arac.,”

Proteins, vol. 74, no. 1, pp. 81–91, 2009.

[21] R. Schleif, “Arac protein, regulation of the l-arabinose operon in escherichia coli, and the light switch mechanism of arac action.,” *FEMS Microbiol Rev*, vol. 34, no. 5, pp. 779–796, 2010.

[22] B. Saviola, R. Seabold, and R. Schleif, “Arm-domain interactions in arac.,” *J. Mol. Biol.*, vol. 278, no. 3, pp. 539–548, 1998.

[23] J. Weldon, M. Rodgers, C. Larkin, and R. Schleif, “Structure and properties of a truely apo form of arac dimerization domain.,” *Proteins*, vol. 66, no. 3, pp. 646–654, 2007.

[24] M. Wu and R. Schleif, "Mapping arm-dna-binding domain interactions in arac.," *J. Mol. Biol.*, vol. 307, no. 4, pp. 1001–1009, 2001.

[25] M. Wu and R. Schleif, "Strengthened arm-dimerization domain interactions in arac.," *J. Biol. Chem.*, vol. 276, no. 4, pp. 2562–2564, 2001.

[26] J. Seedorff and R. Schleif, "Active role of the interdomain linker of arac.," *J. Bacteriol.*, vol. 193, no. 20, pp. 5737–5746, 2011.

[27] F. Malaga, O. Mayberry, D. Park, M. Rodgers, D. Toptygin, and R. Schleif, "A genetic and physical study of the interdomain linker of e. coli arac protein–a trans-subunit communication pathway.," *Proteins*, vol. 84, no. 4, pp. 448–460, 2016.

[28] L. Mayne, "Hydrogen exchange mass spectrometry.," *Methods Enzymol.*, vol. 566, pp. 335–356, 2016.

[29] M. Rodgers and R. Schleif, "Heterodimers reveal that two arabinose molecules are required for the normal arabinose response of arac.," *Biochemistry*, vol. 51, no. 41, pp. 8085–8091, 2012.

[30] A. Tischer, V. Machha, J. Frontroth, M. Brehm, T. Obser, R. Schneppenheim, L. Mayne, S. Walter Englander, and M. Auton, "Enhanced local disorder in a clinically elusive von willebrand factor provokes high-affinity platelet clumping.," *J. Mol. Biol.*, vol. 429, no. 14, pp. 2161–2177, 2017.

[31] Z. Kan, L. Mayne, P. Chetty, and S. Englander, "Exms: data analysis for hx-ms experiments.," *J. Am. Soc. Mass Spectrom.*, vol. 22, no. 11, pp. 1906–1915, 2011.

[32] Z. Kan, B. Walters, L. Mayne, and S. Englander, "Protein hydrogen exchange at residue resolution by proteolytic fragmentation mass spectrometry analysis.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 41, pp. 16438–16443, 2013.

[33] W. Hendrickson and R. Schleif, "Regulation of the escherichia coli l-arabinose operon studied by gel electrophoresis dna binding assay.," *J. Mol. Biol.*, vol. 178, no. 3, pp. 611–628, 1984.

[34] R. Aurora and G. Rose, "Helix capping.," *Protein Sci*, vol. 7, no. 1, pp. 21–38, 1998.

[35] N. R. Kallenbach and Y. Gong, "C-terminal capping motifs in model helical peptides," *Bioorg. Med. Chem.*, vol. 7, no. 1, pp. 143–151, 1999.

[36] M. Auton, J. Rösgen, M. Sinev, L. Holthauzen, and D. Bolen, "Osmolyte effects on protein stability and solubility: a balancing act between backbone and side-chains.," *Biophys Chem*, vol. 159, no. 1, pp. 90–99, 2011.

[37] H. Schott, "Direct comparison of the strength of hydrogen bonds formed by h2o and d2o," *Journal of Macromolecular Science, Part B: Physics*, vol. 27, no. 1, pp. 119–123, 1988.

[38] M. Rodgers and R. Schleif, "Solution structure of the dna binding domain of arac protein.," *Proteins*, vol. 77, no. 1, pp. 202–208, 2009.

[39] N. Finley and P. Rosevear, "Introduction of negative charge mimicking protein kinase c phosphorylation of cardiac troponin i. effects on cardiac troponin c.," *J. Biol. Chem.*, vol. 279, no. 52, pp. 54833–54840, 2004.

[40] S. Dirla, J. Chien, and R. Schleif, "Constitutive mutations in the escherichia coli arac protein.," *J. Bacteriol.*, vol. 191, no. 8, pp. 2668–2674, 2009.

[41] M. Brown, *Matt Brown's Ph.D. Thesis*. PhD thesis, Johns Hopkins University, Baltimore, Maryland, 2019.

[42] A. Brunelle and R. Schleif, "Determining residue-base interactions between arac protein and arai dna.," *J. Mol. Biol.*, vol. 209, no. 4, pp. 607–622, 1989.

[43] E. Pettersen, T. Goddard, C. Huang, G. Couch, D. Greenblatt, E. Meng, and T. Ferrin, "Ucsf chimera–a visualization system for exploratory research and analysis.," *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, 2004.

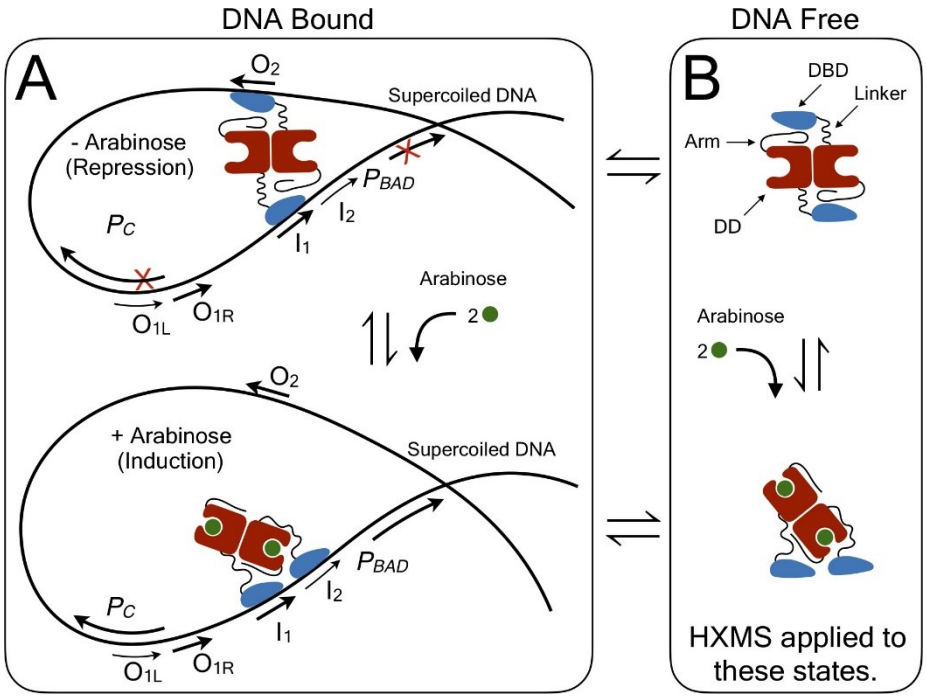

Figure 1: Thermodynamic cycle illustrating arabinose binding to AraC in DNA bound and free states. A) Arabinose-induced light switch mechanism between the DNA looped repressed state and the unlooped activated state. B) Arabinose-induced conformational selection in the absence of DNA. HXMS is applied to these conformational states.

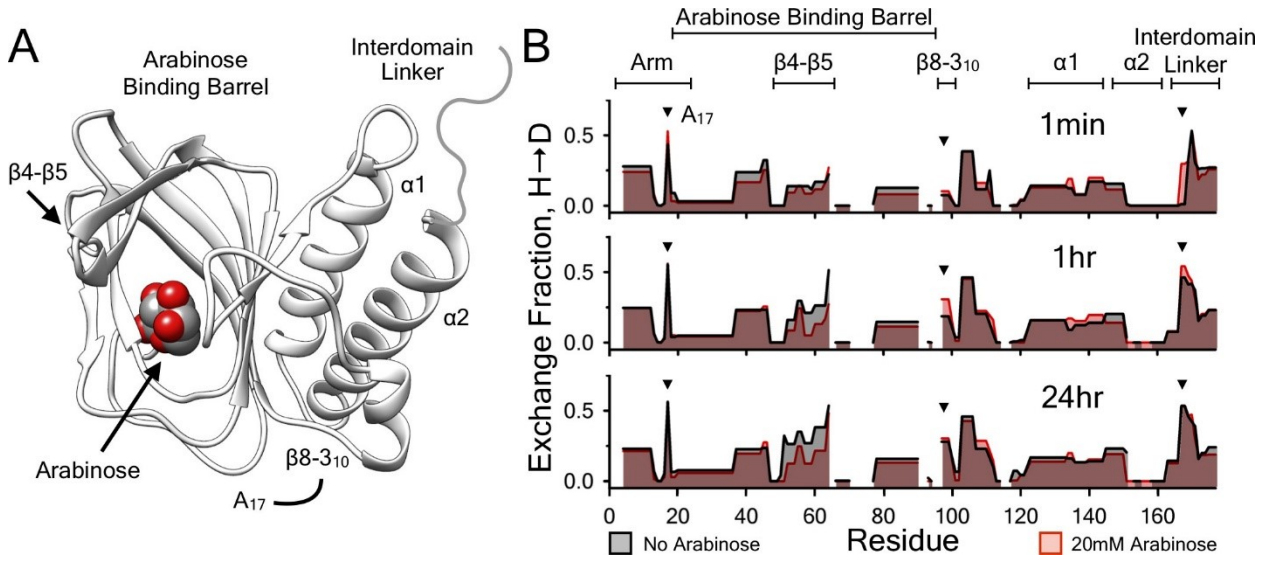

Figure 2: A) Structure of the AraC dimerization domain (pdb ID = 2ARC [8]) highlighting the secondary structures which respond to arabinose binding. Rendered using UCSF Chimera [43]. B) Comparison of the hydrogen exchange of the AraC dimerization domain in the presence (red) and absence (black) of 20mM arabinose as a function of residue number throughout the protein at three exchange time points, 1min (top), 1hr (middle), and 24hrs (bottom). HXMS was quantified in triplicate at each exchange time point. The majority of AraC hydrogen exchange is unchanged upon binding arabinose. Structural regions that show a kinetic structural effect of the binding (see **Figs.3, 4, & 5**) are indicated by horizontal bars. The exchange kinetics are quenched by arrabinose binding in the arabinose binding barrel containing the *β*4-*β*5 hairpin loop. The exchange kinetics are enhanced in the *β*8-3_{10} loop of the arabinose binding barrel connected through space to A_{17} in the N-terminal arm. The exchange kinetics are also enhanced in the allosteric C-terminal interdomain linker containing residuesI_{167}-N_{168}.

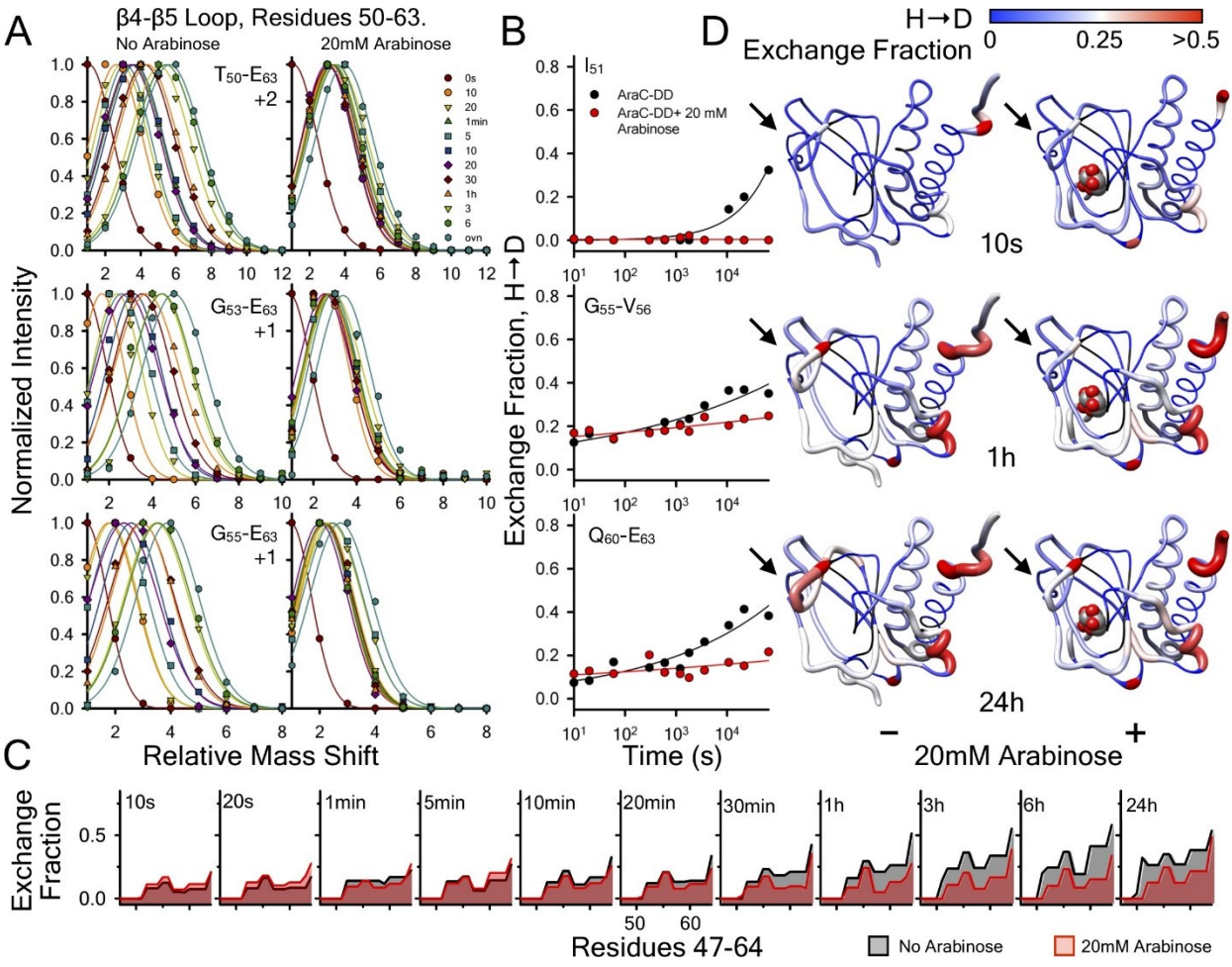

Figure 3: Arabinose binding quenches the hydrogen exchange within the *β*4-*β*5 loop of the dimerization domain near the arabinose binding pocket. HXMS was quantified in triplicate at times between 10s-6h and 24h. **A)** Peptide envelopes (normalized intensity versus mass shift relative to the ”all H” peak) of three peptides covering residues 50-63. Top: T_{50}-E_{63} charge state +2. Middle: G_{53}-E_{63} charge state +1. Bottom: G_{55}-E_{63} charge state +1. Relative to the ”all H” peptide at zero time, HX is observed to continually increase

as a function of time in the absence of arabinose, whereas, the overall HX of these peptides remains fairly constant as a function of time in the presence of arabinose. **B)** Kinetics of HX (exchange fraction versus time) within site-resolved regions of the β4-β5 loop. Top: I_{51}. Middle: G_{55}-V_{56}. Bottom: Q_{60}-E_{63}. **C)** Site resolution of the HX exchange fraction versus residue number. Quenching of HX (in panels B and C) is primarily observed at the longer incubation times between 30min and 24h. **D)** Hydrogen-deuterium exchange fraction mapped onto the structure apo-AraC (pdb ID = 1XJA [23]) and holo-AraC (pdb ID = 2ARC [8]). Black = not resolved, blue = 0, white = 0.25, red 0.5. Rendered using UCSF Chimera [43]. Arrows indicate the structural location of the β4-β5 loop.

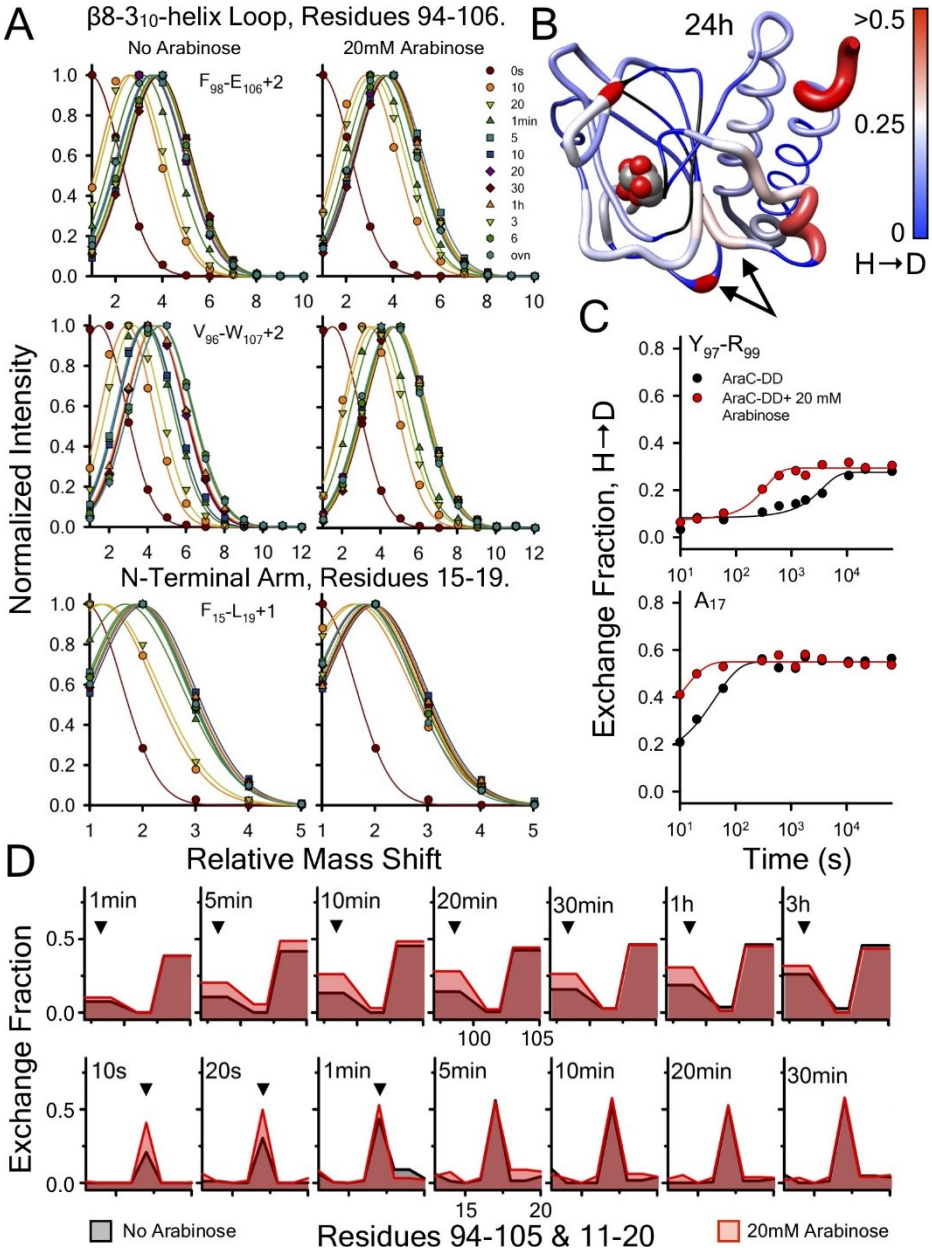

Figure 4: Arabinose binding enhances the hydrogen exchange in residues Y_{97}-R_{99} and A_{17} connected through space in β8-3$_{10}$-helix loop and the N-terminus, respectively. HXMS was quantified in triplicate at times between 10s-6h and 24h. A) Peptide envelopes (normalized intensity versus mass shift relative to the "all H"

peak) of three peptides covering the region. Top: F_{98}-E_{106} charge state +2. Middle: V_{96}-W_{107} charge state +2. Bottom: F_{15}-L_{19} charge state +1. B) Hydrogen-deuterium exchange fraction mapped onto the structure of AraC (pdb ID = 2ARC [8]). Black = not resolved, blue = 0, white = 0.25, red 0.5. Rendered using UCSF Chimera [43]. Arrows indicate the structural location of this region. C) HX Kinetics of site-resolved Y_{97}-R_{99} and A_{17} residues are observed to be faster in the presence of arabinose than in its absence. D) HX kinetics of the region including and surrounding Y_{97}-R_{99} between 1min and 3h and A_{17} between 10s and 30min illustrate localized HX kinetics. Arrow heads indicate the residues involved in the altered kinetics.

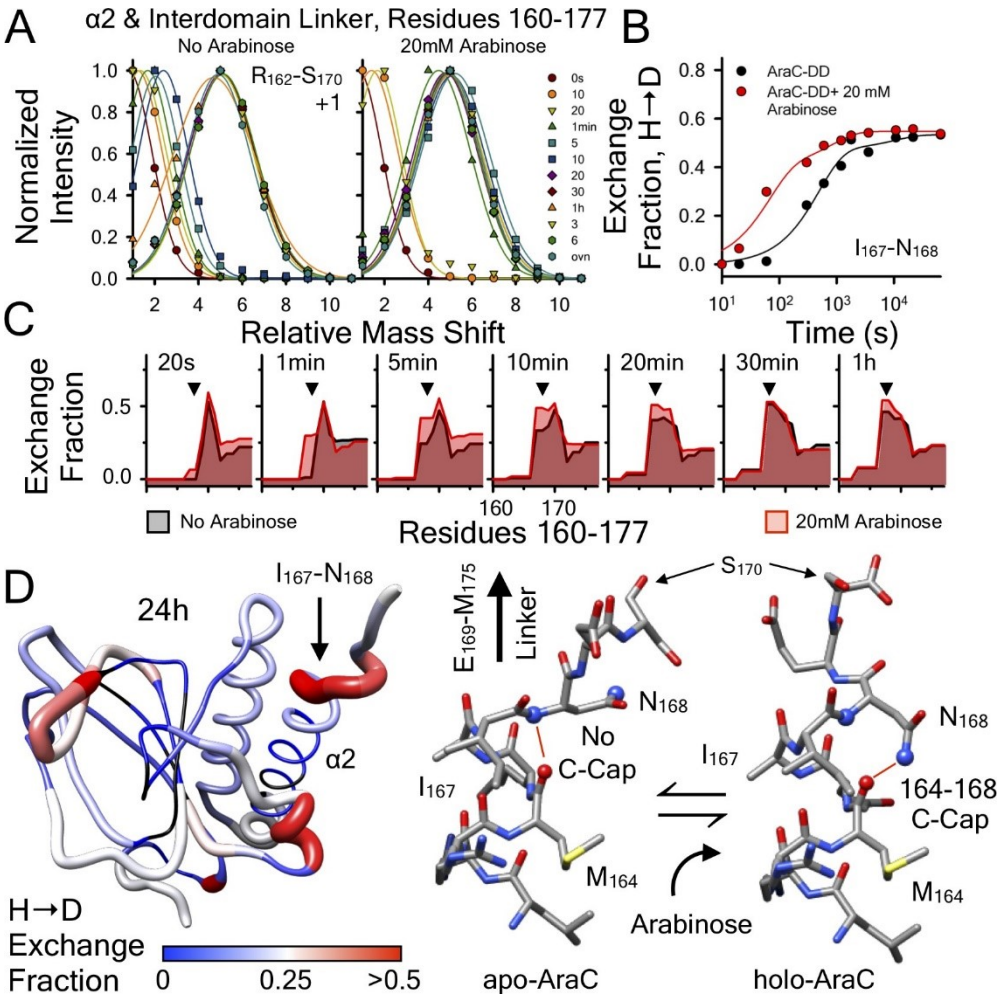

Figure 5: Arabinose binding enhances the hydrogen exchange in residues I_{167}-N_{168} of the C-terminal linker of the dimerization domain. HXMS was quantified in triplicate at times between 10s-6h and 24h. **A)** Peptide envelopes (normalized intensity versus mass shift relative to the "all H" peak) of the R_{161}-S_{169} charge state +1 peptide. **B)** HX Kinetics of site-resolved I_{167}-N_{168} residues of the C-terminal linker are observed to be faster in the presence of arabinose than in its absence. **C)** HX kinetics of the region including and surrounding I_{167}-N_{168} between 20s and 1h illustrate the difference in HX kinetics is localized to I_{167}-N_{168} with only moderate differences in the surrounding residues. Arrow heads indicate the I_{167}-N_{168} residues involved in the altered kinetics. **D)** Hydrogen-deuterium exchange fraction mapped onto the apo structure of AraC (pdb ID = 1XJA [23]) and potential C-terminal Cap equilibrium of the α2 helix containing residues

I_{167}-N_{168}. Structures of apo-AraC (pdb ID = 1XJA [23]) and holo-AraC (pdb ID = 2ARC [8]) C-terminus of the α2 helix interdomain linker junction.

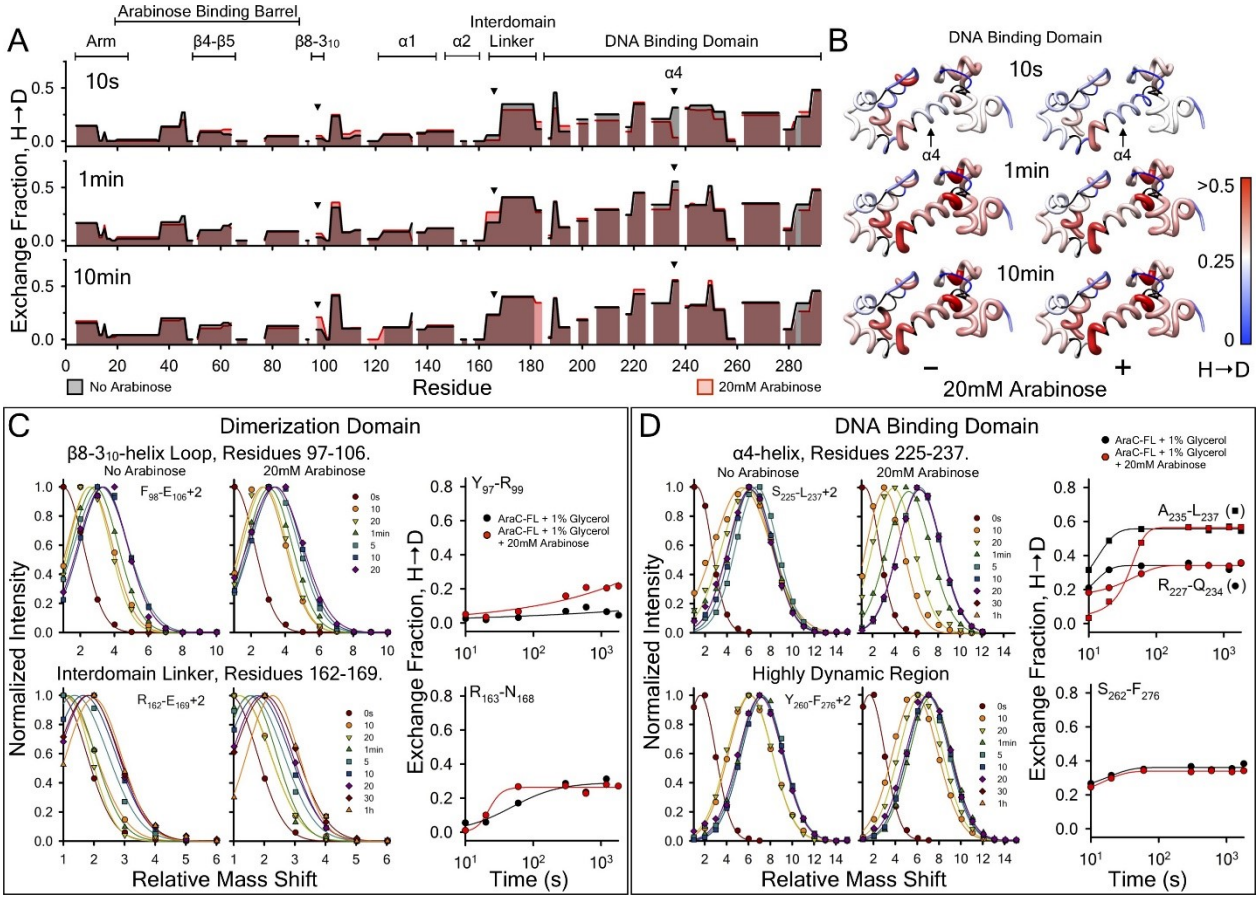

Figure 6: A) Comparison of the hydrogen exchange of the full length AraC protein in the presence (red) and absence (black) of 20mM arabinose as a function of residue number throughout the protein at three exchange time points, 10s (top), 1min (middle), and 10min (bottom). HXMS was quantified in triplicate at each exchange time point. B) Hydrogen-deuterium exchange fraction mapped onto the structure of AraC DNA binding domain (pdb ID = 2K9S [38]). Black = not resolved, blue = 0, white = 0.25, red 0.5. Rendered using UCSF Chimera [43]. C) Peptide envelopes (normalized intensity versus mass shift relative to the "all H" peak) of two peptides covering the dimerization domain in full length AraC which confirm observations in the dimerization domain (**Figs.4 & 5**). Top: F_{98}-E_{106} charge state +2. Bottom: R_{162}-E_{169} charge state +2. HX Kinetics of site-resolved Y_{97}-R_{99} and R_{162}-N_{168} residues are observed to be faster in the presence of arabinose than in its absence. D) Peptide envelopes of two peptides covering the DNA binding domain in full length AraC. Top: S_{225}-L_{237} charge state +2. Bottom: Y_{260}-F_{276} charge state

+2. HX Kinetics of site-resolved residues in α4, A_{235}-L_{237} and R_{227}-N_{234} are observed to be slower in the presence of arabinose than in its absence. S_{262}-L_{276} has fully exchanged in 1min.

Curriculum Vitae

Matthew Brown
Born January 1st, 1984 in Boston Massachusetts

Education

Ph.D. Johns Hopkins University, Program in Cell, Molecular, Developmental Biology and Biophysics

B.S. University of Massachusetts Amherst, Majored in Biology Graduated in May 2010

Laboratory Experience

Graduate Researcher **2012-present**
Laboratory of Robert Schleif at Johns Hopkins University

- Performed biochemical studies to determine the allosteric mechanisms of *Escherichia coli* AraC protein
- Finalized the lab's allosteric model of AraC regulation
- Performed a number of experiments that highlighted the importance of helicity in AraC the interdomain linker
- Mentored a rotation student studying AraC-DNA binding
- Participated in the experimental design and preparation of figures and manuscripts for publication
- Maintained various items of lab equipment
- Managed hazardous waste materials

Laboratory Technologist **2010-2012**
Correlagen Diagnostics, Waltham, MA

- Tested sample DNA obtained from medical patients
- Set up and ran PCR, PCR cleanup, and Sequencing reactions
- Operated and managed Protedyne™ Liquid-handling robots

Undergraduate Research **2010**
Howard Hughes Medical institute Undergraduate Science Program: Gene & Genome Analysis Laboratory
University of Massachusetts Amherst

- Investigated novel gene At1g34300 in *Arabidopsis thaliana*
- Lab techniques used include: DNA & RNA extraction, spectrophotometric &gel- based quantification methods, restriction analysis, PCR, phenotype analysis of T-DNA insertion mutants, microarray analysis, RT-PCR
- Bioinformatics techniques used include: NCBI protein & nucleotide BLAST, Web Map, GENSCAN, GeneSeqer, Primer3, ClustalW, NCBI Gene Expression Omnibus

PUBLICATIONS and PRESENTATIONS

Journal Articles
Brown M, Schleif R. 2019. Helical Behavior of the Interdomain Linker of the Escherichia coli AraC Protein. *Biochemistry* 2019 **58** (26), 2867-2874

Tischer A, **Brown M**, Schleif R, Auton M. 2019. Arabinose Alters Both Local and Distal H–D Exchange Rates in the Escherichia coli AraC Transcriptional Regulator. *Biochemistry* 2019 **58** (26), 2875-2882

Presentations

- Poster presented at the Biophysical Society 63rd Annual Meeting **2018**

"Helicity in the interdomain linker in AraC"

Teaching Experience

Teachers' assistant for Virology at Johns Hopkins University **Spring 2019**
- Graded exams and homework assignments

Teachers' assistant for Chromatin at Johns Hopkins University **Fall 2018**
- Graded exams and homework assignments

Laboratory teachers' assistant for Cell Biology at Johns Hopkins University **Spring 2018**
- Single instructor of a lab section (approx. 20 students) where I communicated the concepts and proper lab techniques to complete various experiments
- Graded lab reports and exams

Laboratory teachers' assistant for Advanced Molecular Biology at Johns Hopkins University **Fall 2017**
- Met with students to aid in understanding subject matter
- Graded homework assignments and exams

Laboratory teachers' assistant for General Biology II at Johns Hopkins University **Spring 2017**
- Single instructor of a lab section (approx. 30 students) where I communicated the concepts and proper lab techniques to complete various experiments
- Graded lab reports and exams

Laboratory teachers' assistant for Biochemistry at Johns Hopkins University **Fall 2016**
- Single instructor of a lab section (approx. 20 students) where I communicated the concepts and proper lab techniques to complete various experiments
- Graded lab reports and exams

Laboratory teachers' assistant for Advanced Molecular Biology at Johns Hopkins University **Spring 2016**
- Met with students to aid in understanding subject matter
- Graded lab reports and exams

Laboratory teachers' assistant for Biochemistry at Johns Hopkins University **Fall 2015**
- Single instructor of a lab section (approx. 20 students) where I communicated the concepts and proper lab techniques to complete various experiments
- Graded lab reports and exams

Laboratory teachers' assistant for Cell Biology at Johns Hopkins University **Spring 2015**
- Single instructor of a lab section (approx. 20 students) where I communicated the concepts and proper lab techniques to complete various experiments
- Graded lab reports and exams

Laboratory teachers' assistant for Biochemistry at Johns Hopkins University **Fall 2014**

- Single instructor of a lab section (approx. 15 students) where I communicated the concepts and proper lab techniques to complete various experiments
- Graded lab reports and exams

Affiliations/Memberships

Biophysical Society **2019**

SKILLS and TECHNIQUES

Laboratory Skills

Assay Development
Protein Biochemistry
Electrophoresis
Spectroscopy
Fluorescence Microscopy
Protein Purification
Molecular Cloning
Protein Kinetics
Sanger DNA Sequencing

Computer Skills

Microsoft Office: Word, Excel, Powerpoint
Adobe: Photoshop, Illustrator
Mathematica
Python
Unix
Pymol